Frequentist Guarantees of Distributed (Non)-Bayesian Inference

Bohan Wu bw2766@columbia.edu
Department of Statistics
Columbia University
New York, NY 10027, USA

César A. Uribe cauribe@rice.edu
Department of Electrical and Computer Engineering
Rice University
Houston, TX 77005, USA

Editor: TBA

Abstract

We establish Frequentist properties, i.e., posterior consistency, asymptotic normality, and posterior contraction rates, for the distributed (non-)Bayes Inference problem for a set of agents connected over a network. These results are motivated by the need to analyze large, decentralized datasets, where distributed (non)-Bayesian inference has become a critical research area across multiple fields, including statistics, machine learning, and economics. Our results show that, under appropriate assumptions on the communication graph, distributed (non)-Bayesian inference retains parametric efficiency while enhancing robustness in uncertainty quantification. We also explore the trade-off between statistical efficiency and communication efficiency by examining how the design and size of the communication graph impact the posterior contraction rate. Furthermore, we extend our analysis to time-varying graphs and apply our results to exponential family models, distributed logistic regression, and decentralized detection models.

Keywords: Distributed inference, Bayesian theory, Bernstein-von Mises, communication efficiency, estimation over networks

1. Introduction

Modern datasets are frequently generated and stored by distributed systems, including social media, sensor networks, blockchain, and cloud-based databases. However, transmission costs make analyzing these datasets on a centralized machine prohibitively expensive and, in some cases, infeasible. To address this challenge, researchers have turned to distributed algorithms that enable decentralized data-driven decision-making under communication constraints (Borkar and Varaiya, 1982; Tsitsiklis and Athans, 1984; Gubner, 1993). In such systems, a set of agents operates within a communication network structure, where each agent can only share information locally with its neighbors. The agents sequentially analyze the data, with each agent performing inference independently and sharing the results through edges defined by the network structure, which may vary over time (Nedić et al., 2017; Uribe et al., 2022b).

Decentralized or distributed Bayesian inference originates in statistics (DeGroot, 1974; Gilardoni and Clayton, 1993). However, it wasn't until the massive advances in computing power in the past decade that the ideas of distributed inference started regaining interest in the statistics community (Uribe et al., 2022a,b). There is a growing line of works on *Distributed Bayesian inference*, which aims to develop scalable and efficient algorithms for posterior computation on large datasets (Jordan et al., 2018). One of the main challenges in this area is to design data-parallel procedures that can handle massive datasets by breaking them into smaller blocks that can be processed independently on individual machines. Much of the current literature focuses on "one-shot" or "embarrassingly parallel" approaches, which involve only one round of communication between local machines and a central node. In these approaches, estimators or posterior samples are obtained in parallel on local machines, communicated to a central node, and combined to form a global estimator or approximation to the posterior distribution.

From the Markov chain Monte Carlo (MCMC) perspective, there have been several developments in parallel MCMC methods for distributed Bayesian inference (Neiswanger et al., 2013; Wang and Dunson, 2013; Minsker et al., 2014; Wang et al., 2015; Rabinovich et al., 2015; Scott et al., 2016; Li et al., 2017; Minsker et al., 2017). These methods draw samples from the subset posterior in parallel agents and combine the samples to obtain an approximation to the posterior measure for the complete data.

From the variational Bayes perspective, algorithms such as stochastic variational inference (Hoffman et al., 2013) have been proposed for distributed Bayesian inference. These algorithms distribute the data across machines, implement the local variational updates in parallel through stochastic gradient descent (SGD), and update the global variational parameters as a weighted average of local optimums. The variational interpretation of the Bayes rule (Walker, 2006) allows the representation between the variational optimization problem and posterior to go both ways.

In parallel to the success in the statistics community, microeconomics has also seen a surge of interest in Distributed Bayesian inference, often referred to as "non-Bayesian social learning." Notable works in this area have focused on its axiomatic foundations (Epstein et al., 2010), conditions for achieving consensus (Acemoglu et al., 2011), various learning rules (Golub and Sadler, 2017), and the effects of information aggregation (Molavi et al., 2018b). This cross-disciplinary interest underscores the broad relevance and applicability of Distributed Bayesian inference techniques. Here, the agents represent individuals seeking to learn about an underlying state of the world θ. Unlike traditional Bayesian approaches, where each agent makes inferences based on the full data, the "non-Bayesian learning" (as economists call it) model captures how individuals make inferences in the presence of other decision-makers, often with limited access to information and interaction with social networks. The Bayesian distributed learning framework offers a promising solution that retains the desirable properties of Bayesian learning, such as ease of uncertainty quantification and flexibility while incorporating a form of information aggregation that aligns with realistic behavioral assumptions. Indeed, the distributed Bayes rule has been shown to reflect reasonable assumptions about individual behavior in society (Jadbabaie et al., 2012). The distributed framework is analytically tractable under certain distributional assumptions and computationally feasible in general. For a more comprehensive literature review, see (Molavi et al., 2018a).

Studies have also investigated distributed Bayesian frameworks for building large-scale decentralized machine-learning systems. These algorithms generally involve an aggregation step consisting of a weighted geometric or arithmetic average of the received beliefs, followed by a Bayesian update using locally available data. In this setup, the set of hypotheses is assumed to be common across the network, but individual observations may come from different distributions. Recent works, including (Lalitha et al., 2014; Qipeng et al., 2015; Shahrampour et al., 2015; Rahimian et al., 2015), proposed variations of this approach and established consistent, geometric, and non-asymptotic convergence rates for a general class of distributed algorithms, covering asymptotic analysis, non-asymptotic bounds, time-varying directed graphs, and adversarial agents, transmission, and node failures.

Distributed Bayesian procedures have attracted substantial interest across disciplines such as electrical engineering, statistics, and economics. Yet, *the broad adoption of these methods in the statistical community has been hindered by a lack of rigorous analysis of their statistical properties.* Moreover, understanding these properties is key to deepening our knowledge of the consensus behavior of agents within varying communication patterns, a topic of interest to the electrical engineering and economics communities. In this paper, we investigate the distributed Bayes procedures that arise from applying the stochastic mirror descent (SMD) algorithm to statistical estimation problems (Uribe et al., 2022a,b). *Our work fills a crucial gap by establishing the Frequentist properties of such distributed Bayes procedures, including posterior consistency, asymptotic normality, and posterior contraction rates.* We also explore the tradeoff between statistical efficiency and communication cost by investigating the relationship between the posterior contraction rate and the structure of the communication graph. Furthermore, we illustrate practical applications of the Bernstein von - Mises results to emphasize their utility in uncertainty quantification. Ultimately, we hope to stimulate further interest in distributed Bayes methods within the statistical community and to establish a solid theoretical foundation for distributed Bayesian inference in fields such as economics and electrical engineering.

The rest of the paper is outlined as follows: Section 2 introduces the distributed Bayesian inference problem from an optimization-centric viewpoint and rigorously defines the distributed Bayes posterior. In Section 3, we outline sufficient conditions for the consistency and asymptotic normality of the distributed Bayes posterior. Section 4 establishes the consistency of distributed Bayes posterior (Theorem 3). Section 5 establishes the Bernstein-von Mises theorems under both correct (Theorem 10) and incorrect model specifications(Theorem 14). Section 6 provides both the abstract and concrete upper bounds on the posterior contraction rate (Theorem 15 and Theorem 18), with an emphasis on model misspecification. Our analysis is extended to time-varying graphs in Section 7, where we establish posterior contraction rates under various communication frequency regimes (Theorem 20). In Section 8, we demonstrate the practical use of our findings by establishing Bernstein-von Mises results for three statistical models, including exponential family (Proposition 24), logistic regression models(Proposition 25), and the distributed detection problem (Proposition 27)- a canonical problem in electrical engineering. The paper concludes in Section 9 with a discussion on future research directions following our findings.

Table 1: Table of Notation

		<u>Functions:</u>
D_{KL}	$\triangleq$	Kullback–Leibler divergence,
L	$\triangleq$	loss function,
$\langle .,. \rangle$	$\triangleq$	L_2 inner product, as in $\langle f,g \rangle = \int_{\mathbb{R}} f(x)g(x)dx$.
		<u>Probability Distributions:</u>
$\mathbb{P}, \mathbb{p}$	$\triangleq$	data generating measure, data generating density,
$\mathbb{P}f(X)$	$\triangleq$	expectation of $f(X)$ when $X \sim \mathbb{P}$, same as $\mathbb{E}_{\mathbb{P}} f(X)$,
Π, π	$\triangleq$	prior measure, prior density.
P, p	$\triangleq$	posterior measure, posterior density,
P_t^j, p_t^j	$\triangleq$	posterior measure, density for agent j at iterant (time) t,
		<u>Others:</u>
G, A, A_{ij}	$\triangleq$	Graph, adjacency matrix and its $(i,j)^{th}$ entry,
$\lambda_j(A)$	$\triangleq$	j^{th} largest eigenvalue of A,
$\mathbf{1}$	$\triangleq$	the vector of all ones.

2. Preliminaries

Suppose we observe a sequence of i.i.d random variables $X_1, X_2, \cdots$ all taking values in a probability space $(\mathcal{X}, \mathbb{B}, \mathbb{P}_0)$ where the true distribution $\mathbb{P}_0$ is unknown. Moreover, let a family of probability distribution be given in the form of $\{\mathbb{P}_\theta : \theta \in \Theta\}$ where $(\Theta, \mathcal{A})$ is a measurable space. Each $\mathbb{P}_\theta$ is a probability measure defined on $(\mathcal{X}, \mathbb{B})$ and the mapping $\theta \to \mathbb{P}_\theta(B)$ is measurable for every $B \in \mathbb{B}$. We refer to $\{\mathbb{P}_\theta : \theta \in \Theta\}$ as the statistical model. The parameter space Θ is typically taken as a subset of the Euclidean or Hilbert space to avoid any measurability issue with $\{\mathbb{P}_\theta : \theta \in \Theta\}$ (Ghosal and Van der Vaart, 2017).

The centralized statistical estimation problem is to fund a subset $\Theta_0 \subseteq \Theta$ such that for $\theta_0 \in \Theta_0$, $\mathbb{P}_{\theta_0}$ is the "closest" to $\mathbb{P}_0$ with respect to a metric. Geometrically, the goal is to find a point in the subset of the probability measures $\{\mathbb{P}_\theta : \theta \in \Theta\}$ closest to $\mathbb{P}_0$ under a given topology. The topology is often defined by divergence on the space of probabilities, such as the Kullback-Leibler (KL) divergence, Rényi divergence, etc. The definition of KL, Rényi divergence, and other divergence functions are reviewed in Section A.1 of the Appendix.

This article uses the KL divergence to define the objective function for statistical estimation problems. This is a natural choice because the KL minimization problem is equivalent to the maximum likelihood estimation in statistics. In short, consider the following estimation problem,

$$\min_{\theta \in \Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta). \tag{2.1}$$

Recall the definition of KL divergence,

$$D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta) = -\mathbb{P}_0 \log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_0}.$$

where $\frac{d\mathbb{p}_\theta(x)}{d\mathbb{p}_0(x)}$ is the Radon-Nikodyn derivative of $\mathbb{P}_\theta$ with respect to $\mathbb{P}_0$.

There are a couple of challenges to solving (2.1). First, $\mathbb{P}_0$ is unknown, and we want to estimate $\mathbb{P}_0$ using the realizations of X. For example, suppose the true distribution $\mathbb{P}_0 = N(0,1)$, and $\mathbb{P}_\theta = \{N(\theta,1); \theta \in [-1,1]\}$ where $N(\mu,\sigma^2)$ is the normal distribution with mean μ and σ^2. Then, θ_0 is estimated to be 0 under the KL divergence. Secondly, first-order stochastic methods such as the gradient descent algorithm are the default approach to solving this optimization problem. But the set Θ might be discrete or non-smooth.

Instead of minimizing over points Θ, we "smooth" the problem by minimizing over the probability distributions on Θ as described next.

Let Δ_Θ be the space of probability density functions defined on Θ. It follows that

$$\min_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta) = \min_{p\in\Delta_\Theta} \int_\Theta p(\theta) D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta) d\theta, \tag{2.2}$$

where Δ_Θ is the (hypothetical) space of all probability distributions over Θ, and $\int_\Theta p(\theta)d\theta = 1$, with $p(\theta) \geq 0$. The problem transforms from choosing the appropriate parameter to choosing the appropriate prior. If θ_0 solves (2.1), $p^* = \delta_{\theta_0}$ solves (2.2), therefore the two problems are equivalent. The advantage is that formulation (2.1) is a continuous optimization problem regardless of the topology in which Θ is defined, i.e., Θ might be a finite discrete set.

There is an equivalent formulation of Problem (2.1) as a minimization problem over the space of probability measures on Θ. Given the linearity of the expectation functional, the Riez representation theorem implies the existence of an inner product $\langle ., . \rangle$ that characterizes the expectation over Θ. We then reformulate the KL-minimization problem in terms of linear stochastic optimization as follows:

$$\min_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta) = \min_{p\in\Delta_\Theta} \mathbb{E}_{\mathbb{P}_0} \langle p, -\log \mathbb{p}_\theta(x) \rangle. \tag{2.3}$$

Note that

$$\begin{aligned}
\min_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta) &= \min_{p\in\Delta_\Theta} \int_\Theta p(\theta) D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta) d\theta \\
&= \min_{p\in\Delta_\Theta} \int_\Theta p(\theta) \int_{\mathcal{X}} \mathbb{p}_0(x) \log \frac{\mathbb{p}_0(x)}{\mathbb{p}_\theta(x)} dx d\theta \\
&= \min_{p\in\Delta_\Theta} \int_{\mathcal{X}} \mathbb{p}_0(x) \int_\Theta p(\theta) \log \frac{\mathbb{p}_0(x)}{\mathbb{p}_\theta(x)} d\theta dx, \qquad \text{by Fubini's theorem} \\
&= \min_{p\in\Delta_\Theta} \int_{\mathcal{X}} \mathbb{p}_0(x) \int_\Theta p(\theta) [\log \mathbb{p}_0(x) - \log \mathbb{p}_\theta(x)] d\theta dx \\
&= \min_{p\in\Delta_\Theta} - \int_{\mathcal{X}} \mathbb{p}_0(x) \int_\Theta p(\theta) \log \mathbb{p}_\theta(x) d\theta dx.
\end{aligned}$$

In this reformulation, Δ_Θ is the probability simplex over Θ, and $\langle p, -\mathbb{p}_\theta(x) \rangle$ represents the inner product between the simplex vector p and the negative log-likelihood $\mathbb{p}_\theta(x)$, taken with respect to the expectation under $\mathbb{P}_0$.

This representation enables efficient computational methods from the stochastic optimization literature, and a proven solution is the stochastic mirror descent (SMD) algorithm (Uribe et al., 2022a,b). Algorithm 6 in (Uribe et al., 2022a,b) iterates through a

KL-regularized optimization problem

$$p_{t+1} = \underset{p \in \Delta_\Theta}{\arg\min} \left\{ -\langle p, \log \mathbb{P}_\theta(x_{t+1}) \rangle + \alpha D_{KL}(p \parallel p_t) \right\}. \tag{2.4}$$

At time t, the agent's goal is to maximize the expected log-likelihood given the belief p_t, i.e., minimizing the divergence. The trade-off between the two behaviors is tuned with the learning rate α.

Equation (2.4) is simultaneously the variational representation of a generalized Bayesian inference. We unpack this as follows:

$$\begin{aligned}
p_{t+1} &= \underset{p \in \Delta_\Theta}{\arg\min} \left\{ -\langle p, \log \mathbb{P}_\theta(x_{t+1}) \rangle + \alpha D_{KL}(p \parallel p_t) \right\} \\
&= \underset{p \in \Delta_\Theta}{\arg\min} \left\{ -\mathbb{E}_p \log \mathbb{P}_\theta(x_{t+1}) + \alpha \mathbb{E}_p \log p(\theta) - \alpha \mathbb{E}_p \log p_t(\theta) \right\} \\
&= \underset{p \in \Delta_\Theta}{\arg\min} \left\{ -\frac{1}{\alpha} \mathbb{E}_p \log \mathbb{P}_\theta(X_{t+1}) + \mathbb{E}_p \log p(\theta) - \mathbb{E}_p \log p_t(\theta) \right\} \\
&= \underset{p \in \Delta_\Theta}{\arg\min} \left\{ \mathbb{E}_p \log \frac{p(\theta)}{(\mathbb{P}_\theta(x_{t+1}))^{\frac{1}{\alpha}} p_t(\theta)} \right\} \\
&= \underset{p \in \Delta_\Theta}{\arg\min} \left\{ D_{KL}(p \parallel \tilde{p}) \right\}, \quad \tilde{p} \propto (\mathbb{P}_\theta(x_{t+1}))^{\frac{1}{\alpha}} p_t(\theta).
\end{aligned}$$

A generalized posterior is a probability distribution $p(\theta \mid X_1, \ldots, X_t) \propto \pi(\theta)\tilde{L}_t(\theta)$, where π is a prior and $\tilde{L}_t$ is a surrogate likelihood (Miller, 2021). From this lens, the solution to stochastic mirror descent is a generalized Bayes posterior:

$$p_{t+1}(\theta) \propto (\mathbb{P}_\theta(x_{t+1}))^{\frac{1}{\alpha}} p_t(\theta). \tag{2.5}$$

The parameter α is simultaneously the step size of the SMD update and a temperature parameter in the generalized Bayes' rule. When α is set to 1, we recover the standard Bayes' theorem.

The result above suggests a fascinating interplay between stochastic optimization and Bayesian inference. While this may seem somewhat unexpected, it aligns well with existing literature that characterizes Bayesian rules and their generalizations as equivalent optimization problems (Knoblauch et al., 2022). The connection between Bayesian methods and Stochastic Mirror Descent (SMD) is particularly fruitful to explore, as it enables us to formulate the distributed Bayesian rule due to the distributed SMD algorithm.

2.1 Distributed Bayesian Inference

In this section, we present a background on the distributed Bayesian inference problem. Consider a set of m independent agents, represented by the index $j = 1, ..., m$. Each agent independently observes a sequence of random variables at discrete time steps $t = 0, 1, 2, ...$ The random variable observed by agent j at time t is denoted as X_t^j. The random process observed by agent j is denoted as X^j. The collection of random variables all agents observe at time t is denoted as X_t. In general, each X^j may be endowed with a different probability space with a common parameter space. For simplicity, we assume that all random variables are defined on the measure space $(\mathcal{X}, \mathbb{B})$, and follow a true probability distribution $\mathcal{P}$ with density p.

Given a common set of parameters Θ, the private parametric statistical model that agent j can access is defined as $\mathcal{P}_\Theta^j = \{\mathbb{P}_\theta^j : \theta \in \Theta\}$. Each model $\mathcal{P}_\Theta^j$ has the same support for all $j \in [m]$.

At time t, the agents interact through an undirected communication graph $G_t = (V, E_t)$, where $V = [m]$ denotes the set of agents. An edge $(j, i) \in E_t$ implies that agent j can communicate with agent i at time t.

The weighted adjacency matrix associated with G_t is denoted by A_t. We assume that A_t is a doubly stochastic matrix obtained by normalizing the matrix A_t', where $[A_t']_{ij} = 1$ if there exists a communication link between agent i and agent j and $[A_t']_{ij} = 0$ otherwise. For undirected graphs, one can always construct a doubly stochastic matrix by normalizing the adjacency matrix, so A_t is guaranteed to exist.

In our framework, agents share information by communicating their beliefs, represented by $p_t^j(\theta)$. This notation represents the posterior distribution of θ as perceived by agent j at time t.

Similar to the centralized case, we formulate distributed statistical inference as a distributed optimization problem:

$$\min_{\theta\in\Theta} \sum_{j=1}^{m} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^j). \tag{2.6}$$

For each agent $j \in [m]$, let π^j represent their initial belief or prior. Therefore, we have $p_0^j = \pi^j$. Similar to the statistical estimation for a single agent, the distributed statistical inference problem admits a reformulation via the distributed stochastic mirror descent algorithm (Uribe et al., 2022a,b). The belief of the agent j at time $t+1$ is obtained through the following mirror descent update.

$$p_{t+1}^j = \underset{p\in\Delta_\Theta}{\arg\min} \left\{ -\langle \log \mathbb{p}_\theta^j(x_{t+1}^j), p\rangle + \sum_{i=1}^{m} [A_t]_{ij} D_{KL}(p \parallel p_t^i) \right\}, \tag{2.7}$$

which is solved via the update

$$p_{t+1}^j(\theta) \propto \mathbb{p}_\theta^j(x_{t+1}^j) \prod_{i=1}^{m} (p_t^i(\theta))^{[A_t]_{ij}}. \tag{2.8}$$

The derivation from (2.7) to (2.8) is identical to that of the case of a single agent (2.4)–(2.5).

Equation (2.8) is aptly referred to as the *distributed Bayes rule*, as it generalizes the Bayesian rule to a distributed setting. This distributed Bayes rule induces a *distributed Bayes posterior* p_{t+1}^j, which might be viewed as the belief of each agent after updating its prior belief based on the new data and information received from its neighbors.

We note that despite its name, the distributed Bayes posterior is not a standard posterior distribution. The classic Bayes rule is a special case of this general distributed Bayes rule, where the communication graph has no edges, denoted as $A_t = I_m$. In this scenario, each agent operates independently, with no influence from others

$$p_{t+1}^j(\theta) \propto \mathbb{p}_\theta(x_{t+1}^j) p_t^j(\theta).$$

In the special case where the communication graph G_t is fully connected, each agent can communicate with every other agent. This leads to the distributed Bayes posterior effectively acting as a weighted Bayesian update rule, with equal weights assigned to all agents.

The distributed Bayes posterior becomes a product of the likelihood of the incoming data and a "tempered" posterior with power $1/m$

$$p_{t+1}^j(\theta) \propto \mathbb{p}_\theta(x_{t+1}^j) \prod_{i=1}^m \prod_{k=1}^t (\mathbb{p}_\theta(x_k^i))^{1/m} \prod_{i=1}^m \pi^i(\theta)^{1/m}.$$

These examples show that distributed Bayesian posterior can adapt to the underlying communication structure, balancing individualistic updates (when there are no communications) and collective updates (when the communication happens according to the network).

Our strategy is to analyze $p_{t+1}^j(\theta)$ using the theory of generalized posterior. In the next Section, we provide the measure-theoretic definition of the distributed Bayes posterior as a generalized Bayes posterior.

2.2 Distributed Bayes Posterior

This section provides a rigorous definition of the distributed Bayes posterior we study throughout this paper.

Define the prior measure Π as follows:

$$\Pi(B) = \frac{\int_B \prod_{j=1}^m \pi^j(\theta)^{1/m} d\theta}{\int_\Theta \prod_{j=1}^m \pi^j(\theta)^{1/m} d\theta}.$$

For each t and j, denote $z_t^j = \int_\Theta \mathbb{p}_\theta^j(x_t^j) \prod_{k=1}^{t-1} \prod_{i=1}^m \mathbb{p}_\theta^i(x_k^i)^{[\prod_{\tau=k}^{t-1} A_\tau]_{ij}} \Pi(d\theta)$ and assume $z_t^j < \infty$.

The *distributed Bayes posterior* is defined as the probability measure P_t^j such that for $B \in \mathbb{B}$,

$$P_t^j(B) = \frac{1}{z_t^j} \int_B \mathbb{p}_\theta^j(x_t^j) \prod_{k=1}^{t-1} \prod_{i=1}^m \mathbb{p}_\theta^i(x_k^i)^{[\prod_{\tau=k}^{t-1} A_\tau]_{ij}} \Pi(d\theta). \tag{2.9}$$

The measure-theoretic definition of distributed Bayes posterior is equivalent to the recursive formulation (2.8). The definition shows that the distributed Bayes posterior is the product of the prior distribution and a surrogate likelihood.

The asymptotic properties of posteriors with surrogate likelihood have been studied in Miller (2021), which establish sufficient assumptions for these posteriors to exhibit posterior concentration, asymptotic normality (Bernstein von-Mises), and asymptotically correct Frequentist coverage. However, the one key difference between the distributed Bayes posterior and generalized posterior defined in (Miller, 2021) is that the surrogate likelihood in Miller (2021) depends solely on the sample size t. In contrast, the surrogate likelihood used in distributed Bayes posterior depends on the sample size t, the statistical model $\mathbb{P}_\Theta^j$, and the communication graph structure G. Therefore, the theoretical analysis of distributed Bayes posterior requires careful adjustments of the standard results for generalized posteriors for new constraints and challenges.

We define the following surrogate loss functions f_t^j, f_t, f on Θ.

$$f_t^j(\theta) = -\frac{1}{t}\log \mathbb{p}_\theta^j(x_t^j) - \frac{1}{t}\sum_{k=1}^{t-1}\sum_{i=1}^{m}\left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} \log \mathbb{p}_\theta^i(X_k^i), \quad j \in [m], \tag{2.10}$$

$$f_t(\theta) = -\frac{1}{mt}\sum_{k=1}^{t}\sum_{i=1}^{m} \log \mathbb{p}_\theta^i(X_k^i), \tag{2.11}$$

$$f(\theta) = -\frac{1}{m}\sum_{i=1}^{m} \mathbb{P}_0 \log \mathbb{p}_\theta^i. \tag{2.12}$$

The function f_t^j corresponds to the surrogate likelihood for the distributed Bayes posterior, f_t corresponds to the empirical mean of $\log p_\theta^j(X)$ and a special case of f_t^j when the adjacency matrix has weights of all $1/m$. Informally, f_t^j and f_t have the same asymptotic properties under mild assumptions, but f_t has much nicer statistical behaviors, so one should expect distributed Bayes posterior to behave like a generalized posterior with likelihood given by f_t.

3. Assumptions

Throughout the paper, we impose the following assumptions:

(i) Θ is an open subset of $\mathbb{R}^p$ with standard Euclidean metric d .

(ii) There exists a unique parameter $\theta_0 \in \Theta$ that minimizes Problem (2.6), that is,

$$\theta_0 = \arg\min_{\theta\in\Theta} \sum_{j=1}^{m} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^j). \tag{3.1}$$

The remaining assumptions fall into four categories: communication graph structure, regularity of private statistical models, prior mass assumptions, and consistent testing assumptions. The latter three are standard in Bayes asymptotics with a well-established history (see Appendix Section A.2). The assumptions on the communication graph are relaxed in Section 7 to allow for temporal dependence.

We assume that the communication graph $G_t, t \geq 1$ is static and undirected, i.e., $G_t = G = (V, E)$.

Assumption 1 (Graph Assumptions) *The communication graph G_t is static and undirected, i.e., $G_t = G = (V, E), t \geq 1$. Moreover, G and the adjacency matrix A satisfy*

(a) A is symmetric and row stochastic.

(b) A has positive diagonal entries, i.e. $a_{ii} > 0$ for all $i \in [m]$.

(c) G is connected, i.e., a directed path exists from any agent i to any agent j.

The assumption of connectedness is standard in the literature on network communication to ensure information flow between agents (Shahrampour et al., 2015; Nedić et al., 2017).

The three assumptions ensure that the Markov chain with transition matrix A is irreducible and aperiodic.

We proceed to establish a result for the convergence of A^t to $\frac{1}{m}\mathbf{1}\mathbf{1}^T$ and an upper bound on the convergence rate of $\|A^t - \frac{1}{m}\mathbf{1}\mathbf{1}^T\|_1$.

Lemma 1 *[Acemoglu et al. (2011); Nedić et al. (2017)] Let Assumption 1 hold. The matrix A satisfies*

$$\limsup_{t\to\infty} \sum_{k=1}^{t}\sum_{j=1}^{m} |[A_{ij}^{t-k}] - \frac{1}{m}| \leq \frac{16m^2\log m}{\nu}, \qquad i \in [m]$$

where ν is the smallest positive entry in A.

Lemma 1 states that given fixed A, the sum $\sum_{k=1}^{t}\sum_{j=1}^{m} |[A_{ij}^{t-k}] - \frac{1}{m}|$ is uniformly bounded in t, and scales in the order of $m^2\log m$. The argument for Lemma 1 relies crucially on the geometric convergence rate of an irreducible, aperiodic Markov chain. See Section B.4 in the Appendix for the proof.

The next group of assumptions concerns the smoothness of the log-likelihood of the private statistical models.

Assumption 2 (Regularity Assumptions) *The regularity assumptions on the statistical model $\mathbb{P}_\theta^j = \{\mathbb{P}_\theta^j, \theta \in \Theta\}$ involves, for every $j \in [m]$ and almost surely $[\mathbb{P}_0]$,*

(a) For every $\theta \in \Theta$, $\mathbb{P}_0|\log \mathbb{p}_\theta^j| < \infty$, i.e. the expectation of $|\log \mathbb{p}_\theta^j(X)|$ is finite under $X \sim \mathbb{P}_0$.

(b) The mapping $\theta \mapsto \log \mathbb{p}_\theta^j$ is convex for every x in a neighborhood of θ_0, and $\nabla f(\theta_0 - \epsilon) < 0 < \nabla f(\theta_0 + \epsilon)$ in every coordinate.

(c) The statistical model $\mathbb{P}_\theta^j$ is differentiable in quadratic mean at θ_0 with nonsingular Fisher information matrix $V_{\theta_0}^j$.

(d) there exists a measurable function s^j with $\mathbb{P}_0[s^j(X)]^2 < \infty$ such that, for every θ_1, θ_2 in a neighborhood of θ_0,

$$|\nabla \log \mathbb{p}_{\theta_1}^j - \nabla \log \mathbb{p}_{\theta_2}^j| \leq s^j(x)\|\theta_1 - \theta_2\|$$

(e) the mapping $\theta \mapsto \log \mathbb{p}_\theta^j$ is twice continuously differentiable for every x in a neighborhood of θ_0, and the Fisher information matrix $V_{\theta_0}^j$ exists and is nonsingular.

(f) the mapping $\theta \mapsto \log \mathbb{p}_\theta^j$ is three-times continuously differentiable for every x in a neighborhood of θ_0, the Fisher information matrix $V_{\theta_0}^j$ exists and is nonsingular, and the third-order partial derivatives uniformly bounded by an integrable function in a neighborhood of θ_0.

The regularity conditions described in Assumption 2 are common prerequisites for the asymptotic analysis of posteriordistributions (Van der Vaart, 2000). The intuition is that the differentiability or convexity of the log-likelihood ensures the consistency of the maximum likelihood estimators, and the quadratic means differentiability enables a valid second-order Taylor expansion around the truth. This allows us to describe the asymptotic properties of the posterior using properties of the MLEs.

The next group of assumptions is typically referred to as the prior mass or prior thickness assumption.

Assumption 3 (Prior Assumptions) *For every $j \in [m]$, the prior distribution Π_j satisfies*

(a) $\Pi_j(\mathcal{U}_\epsilon) > 0$ *for all* $\epsilon > 0$*, where* $\mathcal{U}_\epsilon = \{\theta \in \Theta : \frac{1}{m}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^j) < \epsilon\}$.

(b) the density π_j is continuous and positive at θ_0.

The first assumption states that prior put a sufficient amount of mass in a KL neighborhood of the target distribution $\mathbb{P}_0$. An extra continuity assumption of the density π at θ_0 is sometimes assumed to connect the first assumption to a statement about the neighborhood of θ_0 (Van der Vaart, 2000).

Uniform, consistent testing assumptions enable θ_0 to be distinguished with a sequence of test functions. This assumption ensures that asymptotically negligible mass is placed outside a neighborhood of θ_0.

Assumption 4 (Uniform Consistent Testing) *Let $\Theta \subseteq \mathbb{R}^p$. For each $j \in [m]$,*

(a) For all $\epsilon > 0$, there exists $\delta > 0$ such that $\lim_{t\to\infty} \mathbb{P}_0(\inf_{\|\theta-\theta_0\|>\delta} |f_t^j(\theta) - f_t^j(\theta_0)| \geq \epsilon) = 1$.

(b) There exists a sequence of test functions ϕ_t such that $\mathbb{P}_{\theta_0}^j(1 - \phi_t(X^{(mt)})) \to 0$ *and* $\sup_{\|\theta-\theta_0\|\geq\epsilon} \mathbb{P}_\theta^j(1 - \phi_t(X^{(mt)})) \to 0$ *for every ϵ.*

Assumption 4(b) is attributed to Theorem 10.1 in Van der Vaart (2000), and Assumption 4(a) is given first in Ghosh and Ramamoorthi (2003). The two assumptions in 4 play an identical role in the theory but are slightly different. We mainly use assumptions 4(a).

4. Posterior Consistency

This section establishes the posterior consistency for the distributed Bayes posterior. Theorem 3, to be presented next, provides a general concentration result for the distributed Bayes posterior P_t^j over the measurable space $(\Theta, \mathcal{A})$. The proof of Theorem 3 takes its cues from Schwartz's theorem (Schwartz, 1965). This fundamental theorem implies two supporting findings. Corollary 4 shows that Theorem 3 holds under model misspecification and alternative definitions of the neighborhood around the true parameter. Moreover, Lemma 5 lays down more user - friend sufficient assumptions for the regularity assumption on the log-likelihood (Assumption 2(a)) in Theorem 3.

First, we introduce a supporting lemma to demonstrate that the surrogate loss functions f_t^j and f_t are asymptotically equivalent, and both converge to the population loss function f. This lemma suggests that under the conditions of a connected communication graph and a first-moment condition satisfied by the statistical model, the distributed Bayes posterior P_t^j becomes asymptotically equivalent to a posterior derived from f_t.

Lemma 2 *Let Assumptions 1 and 2(a) hold. Then f_t^j and f_t converges to f on Θ in $[\mathbb{P}_0]-$probability.*

The proof can be bound in Section B.1 of the Appendix. The argument is based on a distributed version of the law of large numbers and geometric convergence of the adjacency matrices described in Lemma 1.

We are ready to state the main result of this section.

Theorem 3 (Consistency) *Let $\theta_0 \in \Theta$ be defined in (3.1), $\epsilon > 0$, and denote $\mathcal{U}_\epsilon = \{\theta \in \Theta : \frac{1}{m}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^j) < \epsilon\}$. Moreover, let Assumptions 1, 2(a), 3(a) hold. Then, the distributed Bayes posterior P_t^j defined in (2.9) has the following property:*

$$\lim_{t\to\infty} P_t^j(\mathcal{U}_\epsilon) = 1,$$

in $[\mathbb{P}_0]-$probability for each $\epsilon > 0$ and $j \in [m]$.

See Section B.1 for the proof. The proof uses the standard technique based on Schwartz's theorem (Miller, 2021). The primary insight is that, given the defined regularity assumptions on private statistical models and the presence of a connected communication graph, the distributed Bayes posterior concentrates at the same point as the $\frac{1}{m}-$geometric average of the individual Bayes posteriors.

Theorem 3 does not assume correct model specification. When the model is misspecified, the distributed Bayes posterior concentrates around the unique minimizer θ_0 of Problem (2.6), which is assumed to exist.

Under additional assumptions, the neighborhoods on the space of probability measures used in stating Theorem 3 can be substituted with neighborhoods on Θ.

Corollary 4 *Let (Θ, d) be metric space, $\theta_0 \in \Theta$, $\epsilon > 0$, and denote $\mathcal{N}_\epsilon = \{\theta \in \Theta : d(\theta, \theta_0) < \epsilon\}$. Moreover, let Assumptions 1, 2(a), 2(b), 3(c) hold. Then the distributed Bayes posterior P_t^j defined in (2.9) has the following property*

$$\lim_{t\to\infty} P_t^j(\mathcal{N}_\epsilon) = 1,$$

in $[\mathbb{P}_0]-$probability for each $\epsilon > 0$ and $j \in [m]$.

The result follows directly from Theorem 3 of Miller (2021). Hence, the proof is omitted.

In practice, Assumption 2(a) is challenging to verify due to the lack of information on the moments of the unknown distribution. The expectation may not exist under a heavy-tailed distribution like the Cauchy distribution. In the result below, we provide an information-theoretic equivalent of Assumption 2(a) that is easier to check.

Lemma 5 *Assumption 2(a) is equivalent to either of the following assumptions:*

(a). $\mathbb{P}_0$ is absolutely continuous with respect to $\mathbb{P}_\theta^j$ for all $\theta \in \Theta$,

(b). $D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^j) < \infty$.

When $\mathbb{P}_0$ has a density $\mathbb{p}_0$, the absolute continuity assumption is equivalent to the assumption that if $\mathbb{p}_\theta^j(x) > 0$, then $\mathbb{p}_0(x) > 0$ for every x (Cover, 1999). The second assumption is the minimum assumption for the problem of distributed statistical estimation to be tractable. Both assumptions are easy to check, as illustrated by the following example.

Example 1 *Assume that $\mathbb{P}_\theta^j = \{N(\theta, \sigma_j^2), \theta \in \Theta\}$ for known $\sigma_j^2 > 0$ and $\mathbb{P}_0 = N(\theta_0, \sigma_0^2)$ for unknown mean θ_0 and variance σ_0^2. The Gaussian distribution with positive variance has support on the whole real line. Therefore, $\mathbb{P}_0$ is absolutely continuous with respect to $\mathbb{P}_\theta^j$ for all j. For the second assumption, the KL divergence between $\mathbb{P}_\theta^j$ and $\mathbb{P}_0$ has an explicit formula: The Kullback-Leibler (KL) divergence between two Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ and $\mathcal{N}(\mu_2, \sigma_2^2)$ is given by:*

$$D_{KL}(\mathbb{P}_\theta^j \parallel \mathbb{P}_0) = \frac{1}{2}\left(\frac{(\theta-\theta_0)^2}{\sigma_0^2} + \frac{\sigma_j^2}{\sigma_0^2} - 1 - \log\frac{\sigma_j^2}{\sigma_0^2}\right).$$

Since $\sigma_j^2, \sigma_0^2, \theta, \theta_0$ are finite, the KL divergence is bounded.

5. Asymptotic Normality

The Bernstein-Von Mises (BvM) theorem states that under certain conditions on the prior, the posterior distribution approximates a Gaussian distribution centered at a consistent estimator, such as the maximum likelihood estimator, as data increases. This theorem is pivotal in Bayesian statistics for at least two reasons. First, it provides a quantitative description of how the posterior contracts to the truth. Second, this theorem justifies Bayesian credible sets as valid Frequentist confidence sets, i.e., sets of posterior probability $1-\alpha$ contain the true parameter at the confidence level $1-\alpha$.

In this section, we establish Bernstein von - Mises theorems for the distributed Bayes posterior defined in Equation (2.9). Our results address the asymptotic normality of the distributed Bayes posterior under both correct and incorrect model specifications. We consider scenarios where all agents accurately specify the model, and some agents gather observations from a 'true distribution' that does not belong to their statistical models.

Theorem 10 provides sufficient assumptions for the distributed posterior to converge to a normal distribution centered around a sequence of M-estimators $\hat{\theta}_t^j$. Theorem 14 generates an analogous result under classical assumptions, with the normal approximation centered around θ_0 (defined in Equation (3.1)). The supporting lemmas are provided preceding the main results. Given the abstract nature of the assumptions, Theorem 10 is supplemented by Corollary 11 and Corollary 12, which outlines more user-friendly assumptions that guarantee the same Bernstein von-Mises (BvM) results.

The Bernstein von - Mises argument relies critically on the following sequence of M-estimators:

$$\hat{\theta}_t^j = \arg\min_{\theta\in\Theta} f_t^j(\theta). \tag{5.1}$$

The existence and consistency of $\hat{\theta}_t^j$ is not always guaranteed; one set of sufficient assumptions is provided as follows.

Lemma 6 *Let Assumptions 1, 2(a), 2(f), 3(b) hold. Then the probability that the equation* $\nabla f_t^j(\hat{\theta}_t^j) = 0$ *has at least one solution converges to* 1 *as* $t \to \infty$*, and there exists a sequence of solutions* $\hat{\theta}_t^j$ *such that* $\hat{\theta}_t^j \xrightarrow{\mathbb{P}_0} \theta_0$*.*

This is a direct consequence of Theorem 5.42 of Van der Vaart (2000); thus, the proof is omitted. Lemma 6 states that the sequence of M estimators exists under third-order smoothness assumptions on the log - log-likelihood. From now on, we assume the existence of M estimators $\hat{\theta}_t^j$ that satisfies (5.1) for every t, j.

The high-order smoothness assumptions on the private log-likelihoods (Assumption (f)) is a strong assumption for guaranteeing consistency. It can be replaced with a more relaxed and amenable convexity assumption (Assumption 2(b)).

Lemma 7 *Let Assumptions 1, 2(a), 2(b) hold. Then* $\hat{\theta}_t^j \xrightarrow{\mathbb{P}_0} \theta_0$*.*

The next two lemmas are useful in the proof of Theorem 10.

Lemma 8 *Let Assumptions 1, 2(c), 4(a) hold and* $\hat{\theta}_t^j \xrightarrow{\mathbb{P}_0} \theta_0$*. Then for every* $\epsilon > 0$*, there exists* $\delta > 0$ *such that*

$$\lim_{t \to \infty} \mathbb{P}_0\Big(\inf_{\|\theta - \hat{\theta}_t^j\| > \delta} |f_t^j(\theta) - f_t^j(\hat{\theta}_t^j)| \geq \epsilon \Big) = 1$$

The result states that if a sequence of estimators $\hat{\theta}_t^j$ is consistent at θ_0, then the existence of uniformly consistent tests at θ_0 (Assumption 4(a)) implies analogous result at $\hat{\theta}_t^j$.

Lemma 9 *Let* $\hat{\theta}_t \in \mathbb{R}^p$ *such that* $\hat{\theta}_t \xrightarrow{\mathbb{P}_0} \theta_0$ *for some* $\theta_0 \in \mathbb{R}^p$*,* π_t *be a density with respect to Lebesgue measure on* $\mathbb{R}^p$*. Suppose* q_t *is the density of* $a_t(\theta - \hat{\theta}_t)$ *where* $\theta \sim \pi_t$ *and* $a_t^{-1} = o(1)$*. If* $\int |q_t(x) - q(x)| dx \to 0$ *in* $[\mathbb{P}_0]-$*probability for some probability density* q*, then for every* $\epsilon > 0$*,*

$$lim_{t \to \infty} \Pi_t(\mathcal{N}_\epsilon) = 1,$$

where N_ϵ *is the neighborhood of* θ_0 *defined in Corollary 4.*

The proofs of Lemma 6, Lemma 7, Lemma 8, and Lemma 9 can be found in Section B.2 of the Appendix. Lemma 6 and 7 relies on the usual consistency argument of M - estimators. Lemma 8 and Lemma 9 use measure-theoretic arguments.

Our first main result of the section provides general sufficient assumptions under which a distributed Bayes posterior exhibits asymptotic normality and an asymptotically correct Laplace approximation, along with the posterior concentration at θ_0.

Theorem 10 (BvM) *Let* Θ *be an open subset of* $\mathbb{R}^p$*. Assume that there exists* $\theta_0 \in \Theta$ *such that* $P_0 = P_{\theta_0}^j$ *for every* $j \in [m]$*. Moreover, let Assumptions 1, 2(a), 2(c), 2(d), 3(b), 4(a) hold. If the sequence* $\hat{\theta}_t^j$ *defined in* (5.1) *satisfies that* $\hat{\theta}_t^j \xrightarrow{\mathbb{P}_0} \theta_0$*, then*

$$f_t^j(\theta) = f_t^j(\hat{\theta}_t^j) - \frac{1}{2}(\theta - \hat{\theta}_t^j)^T \hat{V}_t^j (\theta - \hat{\theta}_t^j) + r_t^j(\theta - \hat{\theta}_t^j), \tag{5.2}$$

where $\hat{V}_t^j$ *is a sequence of matrices that converges in probability to the average Fisher information matrix* $V_{\theta_0} = \frac{1}{m}\sum_{i=1}^m V_{\theta_0}^i$*, and* $|r_t^j(h)| = O(|h|^3)$ *for large enough* t*.*

Let Equation (5.2) and Assumption 3(b), 4(a) hold. As $t \to \infty$, we have

$$\int_{B_\epsilon(\theta_0)} p_t^j(\theta) d\theta \xrightarrow{\mathbb{P}_0} 1 \quad \forall \epsilon > 0. \tag{5.3}$$

that is, the distributed Bayes posterior p_t^j is weakly consistent around θ_0.

Let q_t^j be the density of $\sqrt{t}(\theta - \hat{\theta}_t^j)$ when $\theta \sim P_t^j$. Then,

$$\int_\Theta \left| q_t^j(x) - N(0, V_{\theta_0}^{-1}) \right| dx \xrightarrow{\mathbb{P}_0} 0. \tag{5.4}$$

i.e., the total variational distance between q_t^j and $N(0, V_{\theta_0}^{-1})$ vanishes in $[\mathbb{P}_0]-$probability.

See Section B.2 for the proof. The assumptions of Theorem 10 involve structural and statistical assumptions that are standard within the Bernstein von Mises literature. These include a connected communication graph (Assumption 1), bounded entropy condition of the private statistical models (Assumption 2(a)), a distributed version of the differentiable in quadratic means (DQM) assumption (Assumption 2(c)) and the Lipschitz gradient regularity assumption on the log-likelihood around θ_0 (Assumption 2(d)). The M-estimators, $\hat{\theta}_t^j$, are assumed to be consistent, with sufficient conditions outlined in Lemmas 6 and 7. These regularity assumptions are fairly mild and applicable to, for example, most exponential family models. They serve as the foundation for Equation (5.2), which mirrors the local asymptotic normality assumption in classical BvM theory (Le Cam and Yang, 2000). Additionally, we specify a prior mass assumption (Assumption 3(b)) and an assumption related to uniform, consistent testing (Assumption 4(a)). The prior mass assumption (Assumption 3(b)) is slightly different from the one used in Theorem 3, but they are equivalent when f is continuous at θ_0. These, together with Equation (5.2), are the main sufficient assumptions for the BvM results in Equations (5.3) and (5.4).

The result shows a sampling rate of $\sqrt{t}$, meaning no agent loses parametric efficiency when performing distributed inference in the fully connected network compared to the single-node network. This helps ensure the scalability of the approach. However, the main benefit of communication is the robustness of the uncertainty quantification. In Theorem 10, the normal approximation to the distributed Bayes posterior uses the inverse of the average Fisher information across all agents as a covariance matrix. The intuition is that the individual agent's uncertainty about the ground truth eventually aligns with the network's average level of uncertainty, achieved irrespective of the agent's initial uncertainty or the particular statistical model they employ. Notably, this attribute is a protective mechanism in scenarios where a few agents might be subject to anomalous levels of uncertainty, potentially due to adversarial activities or other forms of data corruption.

The differentiability in quadratic means (DQM) assumption (Assumption 2(c)) in Theorem 10 may be difficult to verify in practical settings. We replace the abstract DQM assumptions with a second-order smoothness assumption on the private log-likelihoods Assumption 2(e).

Corollary 11 *Theorem 10 holds if Assumption 2(c) is replaced with Assumption 2(e).*

Both differentiability in quadratic means (DQM) assumption (Assumption 2(c)) and the Lipschitz gradient assumption (Assumption 2(d)) can be replaced with a third-order smoothness assumption (Assumption 2(f)) which is often called the classical condition for asymptotic normality of M - estimators (Van der Vaart, 2000).

Corollary 12 *Theorem 10 holds if Assumption 2(c) and 2(d) are replaced with Assumption 2(f).*

The proofs of Corollary 11 and Corollary 12 are based on bounding the second and third-order terms in the Taylor expansion of $f_t^j(\theta)$, respectively. We can substitute more model regularity assumptions by bounding higher-order terms in the Taylor expansion.

The Bernstein-von Mises theorem, as traditionally stated (Van der Vaart, 2000), relies on a stochastic rendition of Local Asymptotic Normality (LAN). Here, we introduce a distributed version of stochastic LAN. A distributed family of statistical models $\left(\{\mathbb{P}_\theta^j\}_{j\in[m]}, G\right)$ is *stochastic LAN* at $\theta \in \Theta$ if, for every $j \in [m]$, and in relation to a non-singular scaling factor $\epsilon_t^j \to 0$, there exists a random vector $\Delta_{t,\theta}^j$ and a non-singular matrix V_θ such that $\Delta_{t,\theta}^j$ is bounded in probability and for every compact subset $K \subset \mathbb{R}^p$: as $t \to \infty$,

$$\sup_{h\in K}\left|-tf_t^j(\theta+\epsilon_t^j h)+tf_t^j(\theta)-h^T V_\theta \Delta_{t,\theta}^j+\frac{1}{2}h^T V_\theta h\right| \xrightarrow{\mathbb{P}_0} 0. \tag{5.5}$$

The random vector $\Delta_{t,\theta}^j$ is often called the "local sufficient statistics," and V_θ typically corresponds to the average of Fisher information. It's worth noting that the V_θ matrix must be the same across all agents. The scaling factor ϵ_t^j is chosen to ensure $\Delta_{t,\theta}^j$ is asymptotically normal and V_θ converges to a nonsingular matrix. The default choice is $\epsilon_t^j = \frac{1}{\sqrt{t}}$. However, an agent-specific choice is available if any private statistical model has a different parametric concentration rate.

A sufficient assumption for the distributed LAN is a distributed version of the DQM condition.

Lemma 13 *Let Assumptions 1, 2(a), 2(c) hold. The distributed family of statistical models, denoted as* $\left(\{\mathbb{P}_\Theta^j\}_{j\in[m]}, G\right)$*, is stochastically locally asymptotically normal (LAN) at* θ_0*.*

We now relax the assumption of the correct model specification to explore the more realistic scenario where the data-generating processes are not perfectly aligned with the assumed private statistical models. Specifically, we extend our previous results to the case where the true data-generating distribution $\mathbb{P}_0 \neq \mathbb{P}_{\theta_0}^j$ for some $j \in [m]$.

Theorem 14 (Misspecified BvM) *Let* $\theta_0 \in (\Theta, d)$ *be defined in* (3.1)*. Moreover, let Assumptions 3(b), 4(b) hold, and assume that the distributed stochastic LAN* (5.5) *holds at* θ_0*. Let a sequence of constants* ϵ_t^j *satisfy that for any* $M_t \to \infty$,

$$P_t^j\left(\theta : d(\theta,\theta_0) \geq M_t\epsilon_t^j\right) \xrightarrow{\mathbb{P}_0} 0. \tag{5.6}$$

If q_t^j *is the density of* $(\theta-\theta_0)/\epsilon_t^j$ *when* $\theta \sim P_t^j$*, then*

$$\int_\Theta |q_t^j(x) - N(0, V_{\theta_0}^{-1})|dx \xrightarrow{\mathbb{P}_0} 0,$$

for V_{θ_0} *provided in* (5.5)*.*

Since the covariance matrix V_{θ_0} in the misspecified Bernstein-Von Mises theorem fails to match the average of sandwich covariance matrices, the posterior credible sets derived from P_t^j does not have valid Frequentist coverage. While these sets may be properly centered at $\hat{\theta}_t^j$, their width may be inaccurate, and they don't typically correspond to confidence sets with level $1-\alpha$.

The sequence ϵ_t^j that satisfies assumption (5.6) is called a *posterior contraction rate* of the distributed Bayes posterior P_t^j. This quantity determines the convergence rate of P_t^j to the unknown parameter θ_0. Results to control the contraction rates are provided in the next section.

6. Contraction Rate

Contraction rates quantify the speed at which a posterior distribution approaches the true parameter of the data-generating distribution. Controlling the contraction rates not only refines our understanding of posterior consistency but also helps control the sampling complexity in the misspecified Bernstein von - Mises Theorem (Theorem 14). Unlike the previous sections, which focus on asymptotic results, we provide non-asymptotic bounds and scaling laws of the posterior contraction rates in this section. Our results involve the sample size t, dimension p, and the number of agents m.

For two positive sequences x_t and y_t, we use $x_t \lesssim y_t$ to denote the existence of a constant c, independent of n, such that $x_t \leq c y_t$. Furthermore, we write $x_t \asymp y_t$ when $x_t \lesssim y_t$ and $y_t \lesssim x_t$.

Let (Θ, d) be a metric space. A sequence of constants ϵ_t^j is a posterior contraction rate for P_t^j at the parameter θ_0 if, for every $M_t \to \infty$,

$$P_t^j\left(\theta : d(\theta, \theta_0) \geq M_t \epsilon_t^j\right) \xrightarrow{\mathbb{P}_0} 0, \tag{6.1}$$

The posterior contraction rate is not a unique quantity. Any rate slower than a contraction rate is also a contraction rate. Although the fastest rate is desirable, they may be hard to find. The natural goal is to establish a rate that is a close upper bound for the "optimal" rate. We refer to this upper bound as the contraction rate.

We define a probability measure $\mathbb{P}_\theta$ on the product space $\mathcal{X}^m$. For a measurable set $A \subseteq \mathcal{X}^m$, we have

$$\mathbb{P}_\theta(A) = \frac{\int_A \prod_{j=1}^m [\mathbb{p}_\theta^j(x^j)]^{\frac{1}{m}} dx^1 \ldots dx^m}{\int_\Theta \prod_{j=1}^m [\mathbb{p}_\theta^j(x^j)]^{\frac{1}{m}} dx^1 \ldots dx^m}.$$

The density is given by $\mathbb{p}_\theta(x) = \prod_{j=1}^m [\mathbb{p}_{\theta_0}^j(x^j)]^{\frac{1}{m}}$ for $x \in \mathcal{X}^m$. Given the prior measure Π, we define the *distributed ideal posterior* P_t as the posterior measure corresponding to $\mathbb{P}_\theta$ and Π. For a measurable set $B \subseteq \Theta$, we have

$$P_t(B) = \frac{\int_B \prod_{k=1}^t \mathbb{p}_\theta(x_k) \Pi(d\theta)}{\int_\Theta \prod_{k=1}^t \mathbb{p}_\theta(x_k) \Pi(d\theta)} = \frac{\int_B \prod_{k=1}^t \prod_{j=1}^m [\mathbb{p}_\theta^j(x_k^j)]^{\frac{1}{m}} \Pi(d\theta)}{\int_\Theta \prod_{k=1}^t \prod_{j=1}^m [\mathbb{p}_\theta^j(x_k^j)]^{\frac{1}{m}} \Pi(d\theta)}, \tag{6.2}$$

Let $D_\rho(p \parallel q)$ denote the $\rho-$Rényi divergence, as defined in Section (A.1) of the Appendix. We establish a contraction rate of the distributed Bayes posteriorP_t^j given by:

$$\epsilon_{m,t}^2 + \frac{1}{mt}\mathbb{P}_0 D_{\mathrm{KL}}(P_t^j \parallel P_t) + \frac{1}{m^2 t}\sum_{i=1}^{m} D_{\mathrm{KL}}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^i). \tag{6.3}$$

Here $\epsilon_{m,t}$ is the contraction rate of the distributed ideal posterior P_t. The second term quantifies the approximation error when the distributed Bayes posterior is approximated by the corresponding distributed ideal posterior under the true distribution $\mathbb{P}_0$. The last term captures the minimum average discrepancy between the distributed models and the true data generating distribution.

The main theorem of the section is stated under the "prior mass and testing" framework. The assumptions are sub-exponential refinements of the assumptions for Bernstein von - Mises theorems: (a) The prior is required to put a minimal amount of mass in a neighborhood of the true parameter. (b) Restricted to a subset of the parameter space, there exists a test function that can distinguish the truth from the complement of its neighborhood; (c) The prior is essentially supported on the subset described in (b).

Theorem 15 (Contraction Rate) *Suppose $\epsilon_{m,t}$ is a sequence such that $mt\epsilon_{m,t}^2 \geq 1$. Let C_0, C_1, C_2, $C_3 > 0$ be constants such that $C_0 > C_2 + C_3 + 2$. Let the following assumptions hold:*

1. *For any $\epsilon > \epsilon_{m,t}$, there exists a set $\Theta_t(\epsilon)$ and a testing function ϕ_t such that:*

$$\mathbb{P}_{\theta_0}\phi_t(X^{(mt)}) + \sup_{\substack{\theta \in \Theta_t(\epsilon) \\ d(\theta,\theta_0) \geq C_1\epsilon^2}} \mathbb{P}_\theta(1 - \phi_t(X^{(mt)})) \leq \exp\left(-C_0 t\epsilon^2\right). \tag{C1}$$

2. *For any $\epsilon > \epsilon_{m,t}$, the set $\Theta_t(\epsilon)$ above satisfies:*

$$\Pi\left(\Theta_t(\epsilon)^c\right) \leq \exp\left(-C_0 t\epsilon^2\right). \tag{C2}$$

3. *For some constant $\rho > 1$:*

$$\Pi\left(\theta \in \Theta, \frac{1}{m}\sum_{j=1}^{m} D_\rho(\mathbb{P}_{\theta_0}^j \parallel \mathbb{P}_\theta^j) \leq C_3\epsilon_{m,t}^2\right) \geq \exp\left(-C_2 t\epsilon_{m,t}^2\right). \tag{C3}$$

Then, for the distributed Bayes posterior P_t^j defined in (2.9)*, we have:*

$$\mathbb{P}_0 P_t^j d(\theta,\theta_0) \leq C\left(\epsilon_{m,t}^2 + \gamma_{j,m,t}^2 + \frac{1}{m^2 t}\sum_{j=1}^{m} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j)\right), \tag{6.4}$$

for some constant C depending on C_0, C_1, where the quantity $\gamma_{j,m,t}^2$ is defined as:

$$\gamma_{j,m,t}^2 = \frac{1}{mt}\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t).$$

The proof of Theorem 15 can be found in Section B.2 of the Appendix. It is a consequence of support lemmas that make use of the Gibbs variational representation and the subexponential decay of sub-exponential decay of $d(\theta, \theta_0)$ under P_t.

Assumption (C1) and (C2) is a refinement of assumptions 4 for the uniform consistent testing and states that there is a sequence of tests such that the sum of Type I and Type II errors decrease exponentially with sample size, where the alternative hypothesis is taken in a large enough set under the prior. Assumption (C3) refines the prior mass assumptions 3 by stating that the prior mass decreases exponentially away from a $\rho-$Rényi neighborhood of the true distribution. This assumption is slightly stronger than the equivalent assumption stated with the KL neighborhood because $D_\rho(P \parallel Q) > D_{KL}(P \parallel Q)$ for $\rho > 1$.

The contraction rate is the sum of three terms. The first term $\epsilon_{m,t}^2$ is the contraction rate of the distributed ideal posterior. The second term $\gamma_{j,m,t}^2$ characterizes the distance between the distributed Bayes posterior P_t^j and the ideal posterior P_t. A larger or less connected communication graph means more deviation between the two distributions, which slows the contraction rate. The last term $\frac{1}{m^2 t}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j)$ penalizes the rate by the average discrepancy between the truth and its distributed approximation.

We can show by Markov inequality that the upper bound on $\mathbb{P}_0 P_t^j d(\theta, \theta_0)$ is indeed the contraction rate for the distributed Bayes posterior P_t^j. This allows us to obtain a point estimate $\hat{\theta}$ that converges to the KL minimizer θ_0 at the same rate, for convex loss functions.

Corollary 16 *Under the assumptions of Theorem 15, for any diverging sequence $M_t \to \infty$, we have*

$$\mathbb{P}_0 P_t^j \left(d(\theta, \theta_0) > M_t \left(\epsilon_{m,t}^2 + \gamma_{j,m,t}^2 + \frac{1}{m^2 t} \sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j) \right) \right) \to 0.$$

Furthermore, if $d(\theta, \theta_0)$ is convex in θ, the distributed posterior posterior mean $\hat{\theta} = \int_\Theta \theta dP_t^j(\theta)$ satisfies

$$\mathbb{P}_0 d(\hat{\theta}, \theta_0) \leq C \left(\epsilon_{m,t}^2 + \gamma_{j,m,t}^2 + \frac{1}{m^2 t} \sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j) \right),$$

where C is the same constant in (6.4).

The contraction rate defined in Theorem 15 is somewhat abstract because the terms such as $\epsilon_{m,t}^2$ and $\gamma_{j,m,t}^2$ do not directly inform the design of the underlying communication network. For practical purposes, it's preferable to characterize the rate in terms of design parameters such as m, t, and ν to guide the design of a statistically efficient network structure. To this end, the following section offers more concrete upper bounds on the terms $\epsilon_{m,t}^2$ and $\gamma_{j,m,t}^2$.

The upper bounds for $\epsilon_{m,t}^2$ can be directly borrowed from the existing theory on posterior contraction rates under model misspecification (Kleijn and van der Vaart, 2012). For finite-dimensional models denoted by $\{\mathbb{P}_\theta, \theta \in \Theta\}$, the optimal contraction rate is given by $t^{-1/2}$. However, this result does not follow from Theorem 15 because it requires a more restrictive metric entropy assumption involving the 2^{nd}-order KL divergence. See, for example, Theorem 2.2 of Kleijn and van der Vaart (2012).

The general theory for posterior contraction rates typically combines a prior mass assumption, often in the form of C3, with either a model entropy assumption or a consistent testing assumption in the form of C1 and C2. For a review of the theory of posterior contraction rates, see Chapter 8 of Ghosal and Van der Vaart (2017) and the references therein.

As touched upon in Section 2.2, one should expect the distributed Bayes posterior to be well approximated by the distributed ideal posterior as the sample size increases. The next result provides a uniform bound on the approximation error $\gamma_{j,m,t}^2$.

Lemma 17 *Let Assumptions 1 and 2(a) hold. For P_t^j defined in* (2.9) *and P_t defined in* (6.2)*, we have*

$$\gamma_{j,m,t}^2 \leq \frac{16m\log m}{\nu t}\left(|\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i\in[m]}\inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i)\right)$$

The upper bounds on $\gamma_{j,m,t}^2$ consist of two terms: the first term depends on the structure of the graph, and the second term depends on the graph and the worst-case model misspecification error, denoted by $\max_{i\in[m]}\inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i)$. The intuition is that there is a tradeoff between the size of the communication network and statistical efficiency, and the communication cost could be much higher if one or more agents in the network use misspecified models. If all models are correctly specified, the contraction rate degrades with m at a rate of $m\log m$. If the term $\max_{i\in[m]}\inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i)$ scales at a rate of m^2, then the contraction rate decreases at a much higher rate of $m^3\log m$.

Suppose we ignore the constants and focus on the scaling law. In that case, the contraction rate of the distributed Bayes posterior is a function of the sample size, the number of agents, the smallest positive adjacency weight (spectral gap), the worst-case model misspecification error, and the average model misspecification error. We formalize this in the following result.

Theorem 18 (Practical Contraction Rate) *Let Assumptions 1, 2(a), and the assumptions of Theorem* (15) *hold. Let the distributed ideal posterior P_t satisfies a contraction rate of $\epsilon_{m,t}^2 \lesssim t^{-1}$. For the distributed Bayes posterior P_t^j defined in* (2.9)*, we have:*

$$\mathbb{P}_0 P_t^j d(\theta,\theta_0) \lesssim \frac{1}{t} + \frac{m\log m}{\nu t}\left(1 + \max_{i\in[m]}\inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i)\right) + \frac{1}{m^2 t}\sum_{j=1}^{m} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j).$$

The result follows directly from Theorem 15 and 17. Thus, the proof is omitted.

This theorem outlines how design parameters determine the contraction rate of distributed Bayes posteriors. The second term of the formula is arguably the most interesting: as the number of agents m increases, the contraction rate diminishes proportionally to $m\log m$ and inversely to ν, the smallest positive adjacency weight. This shows that both the scale of the agent network and the minimal communication bandwidth (spectral gap) critically influence the posterior's contraction rate. Additionally, the error scales linearly in the worst-case and average model misspecification errors, which shows that deviations from the ideal model assumptions penalize the contraction rate.

7. Extension to Time-Varying Graphs

This section extends the theory to sequences of time-varying graphs, represented as $G_t = (V, E_t), t \geq 1$. While the node set V remains constant, the edge set E_t undergoes temporal changes. The principal aim of this exploration is to understand the statistical attributes of the distributed Bayes posterior in a fluctuating communication landscape. We focus on one scenario where agents communicate within a fully connected graph with a fixed probability and have no communication otherwise. This setup provides a structured yet flexible framework for analyzing the implications of time-variant connectivity on the statistical efficiency of distributed Bayesian inference.

Assumption 5 *Let Assumption 1 be satisfied for graph G with the adjacency matrix A. We assume that G_t and A_t are independent, random graphs and matrices such that $G_t = G$ and $A_t = A$ with probability $\lambda \in [0, 1]$, respectively, and $A_t = I_m$ otherwise, where I_m is the identity matrix.*

Assumption 5 proposes a setting useful for various reasons. Theoretically, this setting leads to a rigorous study of the tradeoff between statistical efficiency and communication cost. It allows us to explore what constitutes an "optimal" level of communication, given a targeted level of statistical efficiency. On the practical side, this setting can be conceptualized as a network experiencing complete failure with probability λ. In such scenarios, the primary objective is to quantify how this level of intermittent connectivity impacts the overall learning quality within the system.

We provide an analogous result to Lemma 1 in the Assumption 5 setting.

Proposition 19 *Let Assumption 5 hold and $m \geq 2$. Then with probability 1, the sequence of adjacency matrices (A_t) satisfies the following scaling law:*

If $\lambda \geq \frac{2}{m}$, then

$$\limsup_{t\to\infty} \sum_{k=1}^{t} \sum_{j=1}^{m} \left| \left[\prod_{\tau=k}^{t-1} A_\tau \right]_{ij} - \frac{1}{m} \right| \leq \frac{16m^2 \log m + 8m^2 \log \lambda}{\lambda \nu}, \qquad \forall i \in [m]$$

If $0 < \lambda < \frac{2}{m}$, then

$$\limsup_{t\to\infty} \sum_{k=1}^{t} \sum_{j=1}^{m} \left| \left[\prod_{\tau=k}^{t-1} A_\tau \right]_{ij} - \frac{1}{m} \right| \leq \frac{4m^3}{\nu}, \qquad \forall i \in [m]$$

If $\lambda = 0$, then

$$\limsup_{t\to\infty} \sum_{k=1}^{t} \sum_{j=1}^{m} \left| \left[\prod_{\tau=k}^{t-1} A_\tau \right]_{ij} - \frac{1}{m} \right| = \infty, \qquad \forall i \in [m]$$

In these inequalities, ν denotes the smallest positive entry of A.

Proposition 19 provides the scaling laws for three communication regimes. Consensus can be reached as long as the communication occurs with a non-zero probability. However, suppose the goal is to achieve optimal scaling behavior relative to the size of the communication graph (m). Then communication needs to happen with a minimum probability of $\frac{2}{m}$, with a higher frequency being more desirable. On the other hand, if there is a need to adhere to the low-frequency regime ($0 < \lambda < \frac{2}{m}$) due to constraints such as high communication costs, the optimal strategy becomes one of minimal communication. Notably, within this regime, a decrease in communication frequency does not affect the rate of attaining consensus among the agents.

Recall that the contraction rate of P_t^j is decomposed into three pieces.

$$\epsilon_{m,t}^2 + \frac{1}{mt}\mathbb{P}_0 D_{\mathrm{KL}}(P_t^j \parallel P_t) + \frac{1}{m^2 t}\sum_{i=1}^m D_{\mathrm{KL}}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^i).$$

The structure of the communication network affects the contraction rate through the second term. The following Corollary illustrates this.

Corollary 20 *Let Assumptions 5 and 2(a) hold. For P_t^j defined in* (2.9) *and P_t defined in* (6.2)*, the following holds with probability* 1*:*

If $\lambda \geq \frac{2}{m}$, then

$$\frac{1}{mt}\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) \leq \frac{16m\log m + 8m\log\lambda}{\lambda\nu t}\left(|\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i\in[m]}\inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i)\right).$$

If $0 < \lambda < \frac{2}{m}$, then

$$\frac{1}{mt}\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) \leq \frac{4m^2}{\nu t}\left(|\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i\in[m]}\inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i)\right).$$

In Corollary 20, the term "with probability 1" is relative to the measure defined over the sequence $A_1, A_2, \ldots$, where each A_t is an independent random matrix drawn from two deterministic matrices with probability λ. Importantly, this is the same probability measure that underlies Proposition 19, and we continue to use this measure in the sequel.

Varying the communication frequency λ has an effect on the statistical efficiency of performing distributed Bayesian inference over the network. In the following result, we demonstrate the impact of λ on the contraction rates.

Corollary 21 *Let Assumptions 5, 2(a), and the assumptions of Theorem 15 hold.Let $\epsilon_{m,t}^2 \lesssim t^{-1}$. Then for the distributed Bayes posterior P_t^j defined in* (2.9)*, the following holds with probability* 1*:*

If $\lambda \geq \frac{2}{m}$, then

$$\mathbb{P}_0 P_t^j d(\theta, \theta_0) \lesssim \frac{1}{t} + \frac{m\log m + m\log\lambda}{\lambda\nu t}\left(1 + \max_{i\in[m]}\inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i)\right) + \frac{1}{m^2 t}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j).$$

If $0 < \lambda < \frac{2}{m}$, then

$$\mathbb{P}_0 P_t^j d(\theta, \theta_0) \lesssim \frac{1}{t} + \frac{m^2}{\nu t}\left(1 + \max_{i\in[m]}\inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i)\right) + \frac{1}{m^2 t}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j).$$

The distributed ideal posterior P_t does not depend on the communication graph, thus Theorem 15 holds and the assumption that $\epsilon^2_{m,t} \lesssim t^{-1}$ is still justified. Since the only term in the contraction rate that depends on the communication graph is $\gamma^2_{j,m,t}$, the argument for Corollary 21 directly builds on Theorem 15 and Corollary 20. See Section B.3 in the Appendix for the proof.

This is the first result we know of on the impact of time-varying communication networks on the efficiency of distributed statistical inference. The implication deviates from the conventional wisdom about the relationship between communication cost and statistical efficiency. Naturally, one might expect a communication-statistical tradeoff because increasing the communication frequency leads to faster posterior contraction rates. While higher communication frequencies ($\lambda \geq \frac{2}{m}$) do lead to accelerated contraction rates, scaled by a factor of $\frac{\log \lambda}{\lambda}$, the story is different when the frequency is below $\frac{2}{m}$. In this regime, reducing communication costs ceases to have a linear effect on the contraction rate. Instead, the contraction rate scales quadratically with the number of agents m—at a rate of m^2 rather than $m \log m$.

This suggests the possibility of a "phase transition" phenomenon at $\lambda = \frac{2}{m}$, worth exploring for future research. In practical terms, if the goal is to balance communication efficiency and statistical performance in a large network, a communication frequency of $\lambda^* = \frac{2}{m}$ appears to be the optimal choice.

8. Illustrative Examples

8.1 Exponential Family Distributions

Let the distributed statistical models $\left(\{\mathbb{P}^j_\Theta\}_{j\in[m]}, G\right)$ be well - specified, and the Assumption 1 be satisfied for G. Let $\eta : \Theta \to \mathbb{R}^p$ and $T : \mathcal{X} \to \mathbb{R}^p$ be some sufficient statistics and let $\psi^j : \Theta \to \mathbb{R}, h : \mathcal{X} \to \mathbb{R}$ be normalizing functions. We assume that $\mathbb{P}^j_\theta$ is a member of the canonical exponential family

$$\mathbb{p}^j_\theta(x) = h(x) \exp\left(\langle \theta, T^j(x)\rangle - \psi^j(\theta)\right). \tag{8.1}$$

The exponential family includes commonly used Gaussian, exponential, gamma, chi-square, Beta, Dirichlet, Bernoulli, categorical, Poisson, Wishart, inverse Wishart, and geometric distributions. In this section, we only consider the canonical exponential family and interchangeably refer to the exponential and canonical exponential families.

Exponential family distributions are helpful for Bayesian inference because of the conjugate properties, i.e., the posterior is a member of the exponential family if the prior and likelihood are members of the exponential family. This property is preserved in the distributed setting.

Let us consider a scenario where the beliefs of all agents at time t belong to the natural exponential family. In this setting, the belief of agent$i$ concerning the parametersθ can be described as follows:

$$p^i_t(\theta) \propto \exp\left(\langle \theta, \chi^i_t\rangle - B^i(\theta)\right),$$

whereχ^i_t represents the sufficient statistic for agent$i$ at time$t$, and$A^i(\theta)$ is the log-partition function.

We update the one-step-ahead posterior p_{t+1}^j with the distributed Bayes rule,

$$
\begin{aligned}
p_{t+1}^j(\theta) &\propto \log \mathbb{p}_\theta(X_{t+1}^j) \prod_{i=1}^m p_t^i(\theta) \\
&\propto \exp(\langle \theta, T(X_{t+1}^j) \rangle - \psi^j(\theta)) \exp(\langle \theta, \sum_{i=1}^m A_{ij}\chi_t^i \rangle - \sum_{i=1}^m B^i(\theta)) \\
&\propto \exp\left(\langle \theta, T^j(X_{t+1}^j) + \sum_{i=1}^m A_{ij}\chi_t^i \rangle - \psi^j(\theta) - \sum_{i=1}^m B^i(\theta) \right).
\end{aligned}
$$

The distribution p_{t+1}^j is a member of the exponential family with sufficient statistic $T^j(X_{t+1}^j) + \sum_{i=1}^m A_{ij}\chi_t^i$ and the log-partition function $\psi^j(\theta) + \sum_{i=1}^m B^i(\theta)$. This provides an easy-to-implement algorithm to learn the distributed Bayes posterior.

Let the prior be a member of the natural exponential family.

$$
\pi(\theta) = g(u) \exp\left(\langle \theta, u \rangle - \psi^0(\theta) \right). \tag{8.2}
$$

Leveraging the conjugate property of the exponential family, we derive a closed-form expression for the density of the distributed Bayes posterior P_t^j.

Lemma 22 *Let assumption* (1) *hold. Assume that the likelihood* $\mathbb{p}_\theta^i$ *has an exponential family form given by Equation* (8.1)*. Assume that the prior* Π *has an exponential family form given by Equation* (8.2)*. Then the distributed Bayes posterior* P_t^j *defined in* (2.9) *is given by the following formula:*

$$
p_t^j(\theta) = h(X^{(mt)}) \exp(\langle \theta, \chi_t^j + u \rangle - B_t^j(\theta) - \psi^0(\theta)), \tag{8.3}
$$

where

$$
\chi_t^j = \sum_{k=1}^t \sum_{i=1}^m [A_{ji}^{t-k}] T^i(X_k^i), \quad B_t^j(\theta) = \sum_{i=1}^m \sum_{k=1}^t [A_{ji}^{t-k}] \psi^i(\theta), \tag{8.4}
$$

are the sufficient statistic and the log-partition function, respectively.

An exponential family is full-rank if no linear combination of the sufficient statistic is constant. For example, we say the distribution p_θ^i in Equation (8.1) is full-rank if no linear combination of the $D-$dimensional sufficient statistic $T^i(X) = [T_1^i(X), \cdots, T_D^i(X)]$ leads to a constant. This is a mild assumption on the form of the exponential family.

Lemma 23 *Assume that int*$(\Theta) \neq \emptyset$*. Assume that the likelihood* $\mathbb{P}_\theta^j$ *given by Equation* (8.1) *and prior* Π *given by Equation* (8.2) *belong to exponential families with full rank. Then* P_t^j *belongs to an exponential family of full rank. Moreover, the gradient of the log-partition function* ∇B_t^j *is invertible on the interior of* Θ*.*

Lemma 23 establishes the conditions under which the distributed Bayes posterior belongs to the full-rank exponential family. Critically, the invertibility of ∇B_t^j is instrumental for constructing a sequence of consistent M-estimators for θ_0. These estimators serve as the centering sequence for the Laplace approximation, which is key in proving our Bernstein–von Mises result.

Let $\hat{\theta}_t^j$ be the M-estimators corresponding to f_t^j, as defined in Equation (2.10).

$$\hat{\theta}_t^j = \arg\max_{\theta \in \mathbb{R}^p} \langle \theta, \chi_t^j \rangle - B_t^j(\theta).$$

The closed-form expression for $\hat{\theta}_t^j$ can be obtained as follows:

$$\hat{\theta}_t^j = (\nabla_\theta B_t^j)^{-1}(\chi_t^j), \tag{8.5}$$

where χ_t^j, B_t^j are defined in Equation (8.4).

We now state the asymptotic property of the distributed Bayes posteriors for exponential family distributions.

Proposition 24 *Let Θ be an open subset of $\mathbb{R}^p$. Assume that there exists $\theta_0 \in \Theta$ such that $\mathbb{P}_0 = \mathbb{P}_{\theta_0}^j$ for every $j \in [m]$. Moreover, let Assumptions 1, 2(a) hold. Let the sequence of estimators $\hat{\theta}_t^j$ be defined in Equation (8.5). Let q_t^j be the density of $\sqrt{t}(\theta - \hat{\theta}_t^j)$ when $\theta \sim P_t^j$ Then,*

$$\int_\Theta |q_t^j(x) - N(0, V_{\theta_0}^{-1})| dx \xrightarrow{\mathbb{P}_0} 0,$$

where V_{θ_0} is the average of the covariance of T^i evaluated at θ_0, i.e. $V_{\theta_0} = \frac{1}{m}\sum_{i=1}^m Cov(T^i)$.

The BvM result suggests a sample-based Laplace approximation to the distributed Bayes posterior that belongs to the canonical exponential family. One example of such an approximation is the normal distribution with mean given by the moment estimator $\hat{\theta}_t^j$ and covariance given by as$\left(\frac{1}{m}\sum_{i=1}^m \hat{V}(T^i)\right)^{-1}$, where $\hat{V}(T^i)$ is the bootstrapped sample covariance of T^i. By Proposition 24 and Slusky's theorem, we obtain that the total variational distance between P_t^j and the normal distribution $N\left(\hat{\theta}_t^j, \left(\frac{1}{m}\sum_{i=1}^m \hat{V}(T^i)\right)^{-1}\right)$ asymptotically converges to zero.

Laplace approximation enables direct calculation of an asymptotically valid credible region. Let $\chi^2_{\alpha,p}$ be the critical value from a χ^2 distribution with p degrees of freedom. If $\theta \sim N\left(\hat{\theta}_t^j, V_{\theta_0}^{-1}\right)$, a credible region for θ at level $1-\alpha$ is given by the set of θ that satisfies

$$(\hat{\theta}_t^j - \theta)^T \left(\frac{1}{m}\sum_{i=1}^m \hat{V}(T^i)\right)(\hat{\theta}_t^j - \theta)^T \le \frac{\chi^2_{\alpha,p}}{t},$$

where $\hat{V}(T^i)$ can be replaced by any consistent estimator for $\mathrm{Cov}(T^i)$. By Proposition 24 and Slusky's theorem, the credible region has asymptotic coverage of $1-\alpha$ under P_t^j.

8.2 Distributed Logistic Regression

We consider i.i.d. observations $D_k^j = (X_k^j, Y_k^j)$ for $k \in [t]$ and $j \in [m]$ where covariates $X_k^j \in \mathcal{X}$ for $\mathcal{X} \subseteq \mathbb{R}^p$ and responses $Y_k^j \in \{0,1\}$. Let $\theta_0 \in \mathbb{R}^p$ be the true and unknown parameter. A logistic regression model generates the data.

$$Y_k^j \sim \mathrm{Ber}(\nu_k^j), \quad \log\left(\frac{\nu_k^j}{1-\nu_k^j}\right) = \theta_0^T X_k^j. \tag{8.6}$$

For a logistic regression model with coefficient θ, the conditional distribution $\mathbb{p}_\theta(y_k^j \mid x_k^j)$ after marginalizing out ν_k^j is expressed as:

$$\mathbb{p}_\theta(y_k^j \mid x_k^j) = \exp\left(\langle \theta, x_k^j y_k^j \rangle - \sigma(\langle \theta, x_k^j \rangle)\right), \tag{8.7}$$

where $\sigma(\eta) = \log(1+e^\eta)$. The function $\sigma(\eta)$ is strictly convex, since $\sigma''(\eta) = \frac{e^\eta}{(1+e^\eta)^2} > 0$.

Conditioning on the covariates, the model (8.7) aligns with the canonical form of the exponential family. This allows us to apply the results from Section 8.1.

For each agent j, we let the statistical model $\mathcal{P}_\Theta^j$ be a logistic regression model (8.7) with prior $\pi(\theta)$ supported on $\mathbb{R}^p$. By aggregating the log-likelihood, we obtain the closed-form expression for the distributed Bayes posterior:

$$p_t^j(\theta) = \exp\left(\langle \theta, T_t^j \rangle - B_t^j(\theta) - \log \pi(\theta)\right), \tag{8.8}$$

where

$$T_t^j = \sum_{k=1}^{t} \sum_{i=1}^{m} [A_{ji}^{t-k}] X_k^i Y_k^i, \quad B_t^j(\theta) = \sum_{i=1}^{m} \sum_{k=1}^{t} [A_{ji}^{t-k}] \sigma(\langle \theta, X_k^j \rangle).$$

It is difficult to sample from the distribution (8.8) due to the complex form of $\sigma(\theta)$. As a result, the sample-based Laplace approximation characterized by the Bernstein-von Mises theorem becomes especially helpful. Now, we proceed to state this result.

Let $\hat{\theta}_t^j$ be the M-estimators corresponding to f_t^j:

$$\hat{\theta}_t^j = \operatorname*{arg\,max}_{\theta \in \mathbb{R}^p} \langle \theta, T_t^j \rangle - B_t^j(\theta). \tag{8.9}$$

The objective function in (8.9) is strictly concave, guaranteeing the uniqueness of $\hat{\theta}_t^j$. This also enables efficient numerical methods, such as Newton-Raphson, for finding the value of $\hat{\theta}_t^j$.

Proposition 25 *Let $\Theta = \mathbb{R}^p$. Let $D_k^j = (X_k^j, Y_k^j), k \in [t], j \in [m]$ be a sequence of i.i.d. paired random variables generated according to the model* (8.6). *Let Assumptions 1, 2(a) hold. Under these conditions, the estimator sequence $\hat{\theta}_t^j$ defined in Equation* (8.9) *converges to θ_0 in $[\mathbb{P}_0]-$probability.*

Define q_t^j as the density of $\sqrt{t}(\theta - \hat{\theta}_t^j)$, where $\theta \sim P_t^j$. If the following assumptions are satisfied:

i) $\mathbb{P}_{\theta_0} X_1^1 {X_1^1}^T$ exists and is finite and nonsingular.

ii) $\mathbb{P}_{\theta_0} |X_{k,a}^1 X_{k,b}^1 X_{k,c}^1| < \infty$ holds for all $a, b, c \in [p]$, *then we have*

$$\int_{\Theta} \left| q_t^j(x) - N(0, \hat{V}_{\theta_0}^{-1}) \right| \, dx \overset{\mathbb{P}_0}{\to} 0. \tag{8.10}$$

The matrix $\hat{V}_\theta$ is computed as

$$\hat{V}_\theta = \frac{1}{m} \sum_{i=1}^{m} X^{i^T} W^i(\theta) X^i, \tag{8.11}$$

where $W^i(\theta)$ is the diagonal matrix defined by

$$W^i(\theta) = diag\left(\frac{e^{\sum_{j=0}^{p} \theta_j x_{1j}^i}}{\left(1 + e^{\sum_{j=0}^{p} \theta_j x_{1j}^i}\right)^2}, \ldots, \frac{e^{\sum_{j=0}^{p} \theta_j x_{tj}^i}}{\left(1 + e^{\sum_{j=0}^{p} \theta_j x_{tj}^i}\right)^2} \right).$$

The proposition relies on two model-specific assumptions. Assumption i) focuses on the regularity of the private Fisher information. It is a standard prerequisite in Generalized Linear Models (GLMs) to ensure that the model parameter θ is identifiable (Example 16.8, Van der Vaart (2000)). Assumption ii) imposes a more stringent condition requiring a bounded third moment for the covariates. This assumption is used to verify Assumption 2(f) and is generally reasonable in most applications.

Let $\chi^2_{\alpha,p}$ be the critical value from a χ^2 distribution with p degrees of freedom. Proposition 25 specifies an asymptotically valid credible region for the parameter θ under the probability measure P_t^j at level $1-\alpha$. The region is given by the set of θ that satisfies

$$(\hat{\theta}_t^j - \theta)^T \hat{V}_{\theta_0} (\hat{\theta}_t^j - \theta)^T \leq \frac{\chi^2_{\alpha,p}}{t},$$

where $\hat{\theta}_t^j$ is the estimator in Equation (8.9) and $\hat{V}_{\theta_0}$ is defined in Equation (8.11).

Proposition 25 strengthens the robustness of distributed logistic regression models. The averaging Fisher information in the approximate covariance ensures that extreme covariate values do not overly influence the approximate uncertainty. The model-specific assumptions i) and ii) also serve as robustness checks for when the distributed logistic regression model attains the desired asymptotic property. In practice, checking the assumptions and using Laplace approximation provides an efficient, reliable approach to statistical inference in distributed logistic regression models, making it more resilient to various data irregularities.

8.3 Distributed Detection

In this example, we extend the distributed source location scenario from Nedić et al. (2017) to demonstrate our theoretical results in a real-world context. Consider a communication network spread over a unit square $[0,1] \times [0,1]$, comprised of m sensor agents located at points $Z^j, j \in [m]$. The agents are interconnected via a connected undirected graph G, represented by the adjacency matrix A. The goal is to locate a target at an unknown position $\theta_0 \in (0,1) \times (0,1)$. This target generates data $X_t^1, \ldots, X_t^m \sim \delta_{\theta_0}$ at time t, as seen by the agents. However, agents do not receive the data X_t^j directly; instead, the j^{th} agent receives a noisy version of $|X_t^j - Z^j|$, which is the Euclidean distance between the data point X_t^j and the agent's location Z^j.

The goal is for the agents to identify the unknown location parameter, θ_0 collectively. Due to noise and data corruption during transmission, the j^{th} agent employs a statistical model to describe the observed signal $|X_t^j - Z^j|$. Specifically, this signal is modeled as a normal distribution $N(|\theta - Z^j|, {\sigma^j}^2), \sigma^j > 0$ truncated to the interval $[0, |Z^j| + \frac{1}{2}]$.

Let the parameter space be $\Theta = (0,1)^2$. For each $\theta \in \Theta$, the j^{th} agent's statistical model is given by:

$$\mathbb{P}_\theta^j(X_k^j) = \frac{\phi\left(\frac{|X_t^j - Z^j| - |\theta - Z^j|}{\sigma^j}\right)}{\sigma^j\left[\Phi\left(\frac{|Z^j| + \frac{1}{2} - |\theta - Z^j|}{\sigma^j}\right) - \Phi\left(\frac{-|\theta - Z^j|}{\sigma^j}\right)\right]} I(0 \leq |X_t^j - Z^j| \leq |Z^j| + \frac{1}{2}). \tag{8.12}$$

Before collecting any data, the agents collectively decide on a uniform prior for the unknown location parameter θ, defined as:

$$\pi(\theta) = I(\theta \in [0,1]^2). \tag{8.13}$$

The statistical models from Equation (8.12) are based on truncated normal distributions with fixed support, which belongs to the exponential family. The sufficient statistic for $|\theta - Z^j|$ is $|X_t^j - Z^j|$.

We can express the generalized likelihood function $f_t^j(\theta)$, up to a constant, as:

$$\begin{aligned} f_t^j(\theta) = -\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}] \Bigg[& \frac{(|X_t^j - Z^j| - |\theta - Z^j|)^2}{2{\sigma^j}^2} \\ & + \log\left(\Phi\left(\frac{|Z^j| + \frac{1}{2} - |\theta - Z^j|}{\sigma^j}\right) - \Phi\left(\frac{-|\theta - Z^j|}{\sigma^j}\right)\right)\Bigg]. \end{aligned}$$

The M-estimators for the loss function f_t^j are denoted by $\hat{\theta}_t^j$, and they minimize $f_t^j(\theta)$ over the parameter space $[0,1]^2$:

$$\hat{\theta}_t^j = \underset{\theta \in [0,1]^2}{\arg\min} f_t^j(\theta). \tag{8.14}$$

Due to the continuity and strict concavity of the objective function f_t^j, there exists a unique $\hat{\theta}_t^j$, which enables efficient numerical methods like the Newton-Raphson algorithm for its calculation.

Care is needed when f_t^j achieves its minimum at the corners of $[0,1]^2$, namely at the points $(0,0), (0,1), (1,0), (1,1)$. In these scenarios, $\hat{\theta}_t^j$ does not belong to Θ. Including this technicality is primarily to ensure that our definition of $\hat{\theta}_t^j$ satisfies the conditions of the Argmax Theorem (Theorem 3.2.2 in Vaart and Wellner (2023)).

Subsequent results confirm that the sequence of estimators $\hat{\theta}_t^j$ is consistent with θ_0, allowing us to disregard the concern about the extreme corner cases with probability one.

Lemma 26 *Let $\theta_0 \in \Theta$ be given. Under Assumption 2(a), the sequence of estimators $\hat{\theta}_t^j$ defined in Equation* (8.14) *converges to θ_0 in $[\mathbb{P}_0]-$probability. Moreover, $\lim_{t\to\infty} \mathbb{P}_0(\hat{\theta}_t^j \in \Theta) = 1$.*

We can use the Bernstein von - Mises theorem for the distributed location detection problem.

Proposition 27 *Let the private statistical models be given by Equation* (8.12) *and the prior be given by Equation* (8.13). *Let Assumption 1, 2(a) hold. For the sequence of estimators* $\hat{\theta}_t^j$ *defined in Equation* (8.14), *define* q_t^j *as the density of* $\sqrt{t}(\theta - \hat{\theta}_t^j)$, *where* $\theta \sim P_t^j$. *Then we have*

$$\int_\Theta \left| q_t^j(x) - N(0, V_{\theta_0}^{-1}) \right| dx \overset{\mathbb{P}_0}{\to} 0, \tag{8.15}$$

where V_{θ_0} *is given by*

$$V_{\theta_0} = \frac{1}{m} \sum_{j=1}^{m} \frac{(\theta_0 - Z^j)(\theta_0 - Z^j)^T}{{\sigma^j}^4 |\theta_0 - Z^j|^2} \left[\frac{\phi(\frac{|Z^j| + \frac{1}{2} - |\theta_0 - Z^j|}{\sigma^j}) - \phi(\frac{-|\theta_0 - Z^j|}{\sigma^j})}{\Phi(\frac{|Z^j| + \frac{1}{2} - |\theta_0 - Z^j|}{\sigma^j}) - \Phi(\frac{-|\theta_0 - Z^j|}{\sigma^j})} \right]^2. \tag{8.16}$$

Let $\chi^2_{\alpha,p}$ denote the critical value from a chi-squared distribution with p degrees of freedom. Proposition 27 delineates an asymptotically valid credible region for the parameter θ under the probability measure P_t^j at a confidence level of $1 - \alpha$. Specifically, this credible region encompasses all values of θ that satisfy the following inequality:

$$(\hat{\theta}_t^j - \theta)^\top \hat{V}_{\theta_0} (\hat{\theta}_t^j - \theta) \leq \frac{\chi^2_{\alpha,p}}{t},$$

where $\hat{\theta}_t^j$ represents the estimator defined in Equation (8.9). Furthermore, $\hat{V}_{\theta_0}$ serves as the empirical analogue of the covariance matrix V_{θ_0}, as defined in Equation (8.16), with the true parameter θ_0 replaced by its estimate $\hat{\theta}_t^j$.

9. Discussion and Future Directions

We have studied the Frequentist statistical properties of the distributed (Non-Bayesian) Bayesian inference in terms of "posterior" consistency, Bernstein von - Mises theorems, and "posterior" contraction rates. Our results provide the first rigorous insights into the statistical efficiency of distributed Bayesian inference and its dependence on the design parameters of the underlying communication network, such as the number of agents and the network topology. The promising results offer several avenues for future research.

Future work should investigate the Frequentist statistical properties of the distributed Bayes posterior under more complex time-varying network structures, such as networks with single link failures (Shahrampour et al., 2015) or under other random graph structures such as Erdős–Rényi graphs (Erdős et al., 1960). Understanding how distributed Bayesian inference adapts to such scenarios will be crucial for applying theoretical insights in unpredictable or adversarial settings.

Another compelling direction is to explore the Frequentist coverage properties of the distributed Bayes posterior. This could include analyses of confidence intervals or hypothesis tests when the underlying model is well-specified and misspecified. Understanding how the distributed Bayes posterior aligns with Frequentist criteria can offer additional validation and robustness checks for the model.

There is also potential to connect our work with the theory of non-Bayesian social learning in economics, such as the results in Molavi et al. (2018a). Insights from non-Bayesian frameworks like variants of the DeGroot model (DeGroot, 1974; Acemoglu et al., 2011) could shed light on the convergence properties of distributed Bayesian methods, especially in settings where agents have heterogeneous prior beliefs or are influenced by external signals.

Our work finds resonance in the literature on Frequentist distributed inference that focuses on communication efficiency. Future studies could integrate insights from this body of work, including communication-efficient methods (Jordan et al., 2018) and high-dimensional distributed statistical inference (Battey et al., 2015), to design and analyze distributed Bayesian methods.

Lastly, our theoretical findings could inform the design of communication networks in practical applications. Specifically, an analytical understanding of how the number of agents and communication costs interact could guide the construction of more efficient and robust distributed Bayesian systems.

References

Daron Acemoglu, Munther A Dahleh, Ilan Lobel, and Asuman Ozdaglar. Bayesian learning in social networks. *The Review of Economic Studies*, 78(4):1201–1236, 2011.

Heather Battey, Jianqing Fan, Han Liu, Junwei Lu, and Ziwei Zhu. Distributed estimation and inference with statistical guarantees. *arXiv preprint arXiv:1509.05457*, 2015.

Peter J Bickel and Bas J K Kleijn. The semiparametric bernstein–von mises theorem. *The Annals of Statistics*, 40:206–237, 2012. ISSN 0090-5364.

Peter J Bickel and Joseph A Yahav. Some contributions to the asymptotic theory of bayes solutions. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 11:257–276, 1969. ISSN 1432-2064.

Natalia A Bochkina and Peter J Green. The bernstein–von mises theorem and nonregular models. *The Annals of Statistics*, 42(5):1850–1878, 2014.

Vivek Borkar and Pravin Varaiya. Asymptotic agreement in distributed estimation. *IEEE transactions on automatic control*, 27(3):650–655, 1982.

Ismaël Castillo and Judith Rousseau. A bernstein–von mises theorem for smooth functionals in semiparametric models. *The Annals of Statistics*, 43:2353–2383, 2015. ISSN 0090-5364.

Thomas M Cover. *Elements of information theory.* John Wiley & Sons, 1999.

Morris H DeGroot. Reaching a consensus. *Journal of the American Statistical association*, 69(345):118–121, 1974.

Larry G Epstein, Jawwad Noor, and Alvaro Sandroni. Non-bayesian learning. *The BE Journal of Theoretical Economics*, 10:1–20, 2010.

Paul Erdős, Alfréd Rényi, et al. On the evolution of random graphs. *Publ. math. inst. hung. acad. sci*, 5(1):17–60, 1960.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Jayanta K. Ghosh and R.V. Ramamoorthi. *Bayesian Nonparametrics*. Springer-Verlag, 2003. doi: 10.1007/b97842. URL `https://doi.org/10.1007%2Fb97842`.

Gustavo L Gilardoni and Murray K Clayton. On reaching a consensus using degroot's iterative pooling. *The Annals of Statistics*, 21(1):391–401, 1993.

Benjamin Golub and Evan Sadler. Learning in social networks. *Available at SSRN 2919146*, 2017.

John A Gubner. Distributed estimation and quantization. *IEEE Transactions on Information Theory*, 39(4):1456–1459, 1993.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 2013.

Ali Jadbabaie, Pooya Molavi, Alvaro Sandroni, and Alireza Tahbaz-Salehi. Non-bayesian social learning. *Games and Economic Behavior*, 76:210–225, 2012. ISSN 0899-8256.

Michael I Jordan, Jason D Lee, and Yun Yang. Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 2018.

Anya Katsevich. Improved scaling with dimension in the bernstein-von mises theorem for two statistical models. *arXiv preprint arXiv:2308.06899*, 2023.

Bas J K Kleijn and Aad W van der Vaart. The bernstein-von-mises theorem under misspecification. *Electronic Journal of Statistics*, 6:354–381, 2012. ISSN 1935-7524.

Jeremias Knoblauch, Jack Jewson, and Theodoros Damoulas. An optimization-centric view on bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109, 2022.

Anusha Lalitha, Anand Sarwate, and Tara Javidi. Social learning and distributed hypothesis testing. In *2014 IEEE International Symposium on Information Theory*, pages 551–555. IEEE, 2014.

Pierre-Simon Laplace. Supplément au mémoire sur les approximations des formules qui sont fonction de très grands nombres. *Mémoires de l'Académie Royale des sciences de Paris*, 1810:559–565, 1809.

Lucien Le Cam. On some asymptotic properties of maximum likelihood estimates and related bayes estimates. *Univ. California Pub. Statist.*, 1:277–330, 1953.

Lucien Le Cam and Grace Lo Yang. *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2000. ISBN 0387950362.

Cheng Li, Sanvesh Srivastava, and David B Dunson. Simple, scalable and accurate posterior interval estimation. *Biometrika*, 104:665–680, 2017. ISSN 0006-3444.

Marco Avella Medina, José Luis Montiel Olea, Cynthia Rush, and Amilcar Velez. On the robustness to misspecification of α-posteriors and their variational approximations. *Journal of Machine Learning Research*, 23(147):1–51, 2022.

Jeffrey W Miller. Asymptotic normality, concentration, and coverage of generalized posteriors. *J. Mach. Learn. Res.*, 22:168–1, 2021.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David Dunson. Scalable and robust bayesian inference via the median posterior. In *International conference on machine learning*, pages 1656–1664. PMLR, 2014.

Stanislav Minsker, Sanvesh Srivastava, Lizhen Lin, and David B Dunson. Robust and scalable bayes via a median of subset posterior measures. *The Journal of Machine Learning Research*, 18:4488–4527, 2017. ISSN 1532-4435.

Pooya Molavi, Alireza Tahbaz-Salehi, and Ali Jadbabaie. A theory of non-bayesian social learning. *Econometrica*, 86(2):445–490, 2018a.

Pooya Molavi, Alireza Tahbaz-Salehi, and Ali Jadbabaie. A theory of non-bayesian social learning. *Econometrica*, 86:445–490, 2018b. ISSN 0012-9682.

Angelia Nedić, Alex Olshevsky, and César A Uribe. Fast convergence rates for distributed non-bayesian learning. *IEEE Transactions on Automatic Control*, 62:5538–5553, 2017. ISSN 0018-9286.

Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel mcmc. *arXiv preprint arXiv:1311.4780*, 2013.

Maxim Panov and Vladimir Spokoiny. Finite sample bernstein–von mises theorem for semiparametric problems. *Bayesian Analysis*, 10(3):665–710, 2015.

Liu Qipeng, Zhao Jiuhua, and Wang Xiaofan. Distributed detection via bayesian updates and consensus. In *2015 34th Chinese Control Conference (CCC)*, pages 6992–6997. IEEE, 2015.

Maxim Rabinovich, Elaine Angelino, and Michael I Jordan. Variational consensus monte carlo. *Advances in Neural Information Processing Systems*, 28, 2015.

Mohammad Amin Rahimian, Shahin Shahrampour, and Ali Jadbabaie. Learning without recall by random walks on directed graphs. In *2015 54th IEEE Conference on Decision and Control (CDC)*, pages 5538–5543. IEEE, 2015.

Jeffrey S Rosenthal. Convergence rates for markov chains. *Siam Review*, 37(3):387–405, 1995.

Lorraine Schwartz. On bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 4:10–26, 1965. ISSN 1432-2064.

Steven L Scott, Alexander W Blocker, Fernando V Bonassi, Hugh A Chipman, Edward I George, and Robert E McCulloch. Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management*, 11:78–88, 2016. ISSN 1750-9653.

Shahin Shahrampour, Alexander Rakhlin, and Ali Jadbabaie. Distributed detection: Finite-time analysis and impact of network topology. *IEEE Transactions on Automatic Control*, 61:3256–3268, 2015. ISSN 0018-9286.

Xiaotong Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. *Journal of the American Statistical Association*, 97:222–235, 2002. ISSN 0162-1459.

John Tsitsiklis and Michael Athans. Convergence and asymptotic agreement in distributed decision problems. *IEEE Transactions on Automatic Control*, 29(1):42–50, 1984.

Cesar A Uribe, Alexander Olshevsky, and Angelia Nedich. Non-asymptotic concentration rates in cooperative learning part i: Variational non-bayesian social learning. *IEEE Transactions on Control of Network Systems*, 2022a. ISSN 2325-5870.

Cesar A Uribe, Alexander Olshevsky, and Angelia Nedich. Non-asymptotic concentration rates in cooperative learning part ii: Inference on compact hypothesis sets. *IEEE Transactions on Control of Network Systems*, 2022b. ISSN 2325-5870.

AW van der Vaart and Jon A Wellner. Empirical processes. In *Weak Convergence and Empirical Processes: With Applications to Statistics*, pages 127–384. Springer, 2023.

Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3. Cambridge University Press, 2000. ISBN 0521784506.

Stephen G Walker. Bayesian inference via a minimization rule. *Sankhyā: The Indian Journal of Statistics*, pages 542–553, 2006. ISSN 0972-7671.

Xiangyu Wang and David B Dunson. Parallelizing mcmc via weierstrass sampler. *arXiv preprint arXiv:1312.4605*, 2013.

Xiangyu Wang, Fangjian Guo, Katherine A Heller, and David B Dunson. Parallelizing mcmc with random partition trees. *Advances in neural information processing systems*, 28, 2015.

Yixin Wang and David M Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

Appendix A. Review of concepts

A.1 Distances and Divergences

This section reviews a class of divergence functions.

Definition 28 (Rényi divergence) *Let $\rho > 0$ and $\rho \neq 1$. The ρ-Rényi divergence between two probability measures P_1 and P_2 is defined as*

$$D_\rho(P_1 \parallel P_2) = \begin{cases} \frac{1}{\rho-1} \log \int \left(\frac{dP_1}{dP_2}\right)^{\rho-1} dP_1 & \text{if } P_1 \ll P_2, \\ +\infty & \text{otherwise.} \end{cases}$$

The relations between the Rényi divergence and other divergence functions are summarized below:

1. *When $\rho \to 1$, the Rényi divergence converges to the Kullback–Leibler divergence, defined as*

$$D_{KL}(P_1 \parallel P_2) = \begin{cases} \int \log \left(\frac{dP_1}{dP_2}\right) dP_1 & \text{if } P_1 \ll P_2, \\ +\infty & \text{otherwise.} \end{cases}$$

2. *When $\rho = 1/2$, the Rényi divergence is related to the Hellinger distance by*

$$D_{1/2}(P_1 \parallel P_2) = -2 \log \left(1 - H(P_1, P_2)\right),$$

and the Hellinger distance is defined as

$$H(P_1, P_2) = \sqrt{\frac{1}{2} \int \left(\sqrt{dP_1} - \sqrt{dP_2}\right)^2}.$$

3. *When $\rho = 2$, the Rényi divergence is related to the χ^2-divergence by*

$$D_2(P_1 \parallel P_2) = \log \left(1 + \chi^2(P_1 \parallel P_2)\right),$$

and the χ^2-divergence is defined as

$$\chi^2(P_1 \parallel P_2) = \int \left(\frac{(dP_1)^2}{dP_2} - 1\right).$$

Definition 29 (Total variation) *The total variation distance between two probability measures P_1 and P_2 is defined as*

$$TV(P_1, P_2) = \frac{1}{2} \int |dP_1 - dP_2|.$$

A.2 An (incomplete) Summary of Previous BvM Results

The history of the Bernstein von Mises (BvM) theorem is marked by steady evolution, expanding its reach to wider contexts with weaker assumptions. The origins of the Berstein von-Mises (BvM) theorem trace back to Laplace (1809). Earlier works on BvM theorem are restricted to well-specified, i.i.d statistical models with fixed, finite-dimensional parameters and use assumptions that involve up to fourth-order derivatives of the log-likelihood (Le Cam, 1953; Bickel and Yahav, 1969; Le Cam and Yang, 2000)).

This classical Bernstein von - Mises theorem for i.i.d., finite-dimensional data was brilliantly summarized in (Van der Vaart, 2000, Chapter 10). The Theorem relies on three assumptions: the existence of a sequence of consistent estimators, local asymptotic normality (LAN), and uniform consistent testing. One typical sufficient assumption for LAN is differentiable in quadratic means (DQM), which involves only first-order derivatives of the log-likelihood to achieve a desirable quadratic expansion. In addition to the LAN assumption, Schwartz Schwartz (1965) proposed the concept of uniformly consistent test assumptions around the true parameter θ_0.

The contemporary wave of research has extended BvM theorems beyond the canonical settings. These studies have broached areas such as model misspecification (Kleijn and van der Vaart, 2012), semi-parametric models Shen (2002); Bickel and Kleijn (2012); Panov and Spokoiny (2015); Castillo and Rousseau (2015), and parameters positioned on the boundary of the parameter space (Bochkina and Green, 2014). Among these explorations, a particularly relevant line of inquiry has focused on BvM for non-standard Bayes procedures (Knoblauch et al., 2022), including generalized posteriors (Miller, 2021), Bayes rules derived from optimization perspectives, and variational Bayes posteriors (Wang and Blei, 2019; Medina et al., 2022). Another emerging thread of work establishes BvM theorems for high dimensional models, with most recent results allowing the dimension to grow at the order $d^2 \lesssim n$ (Panov and Spokoiny, 2015; Katsevich, 2023).

Appendix B. Proofs

B.1 Proofs of Results in Section 4

Proof [Lemma 2] Define vector - valued functions $\phi_t(\theta) = \left[f_t^1(\theta), \cdots, f_t^m(\theta)\right]^T$ and $\log \mathbb{p}_\theta(X_t) = \left[\log \mathbb{P}_\theta^1(X_t^1), \cdots, \log \mathbb{P}_\theta^m(X_t^m)\right]^T$.
Since $\mathbb{P}_0|\log \mathbb{P}_\theta^i| = \mathbb{P}_0(\log \mathbb{P}_\theta^i)_+ + \mathbb{P}_0(\log \mathbb{P}_\theta^i)_- < \infty$, we have

$$f_t^j(\theta) = -\frac{1}{t}\sum_{k=1}^t\sum_{i=1}^m [A_{ji}^{t-k}] \log \mathbb{p}_\theta^i(X_t^i) = \frac{1}{t}\sum_{k=1}^t \left\langle [A_{j,\cdot}^{t-k}], \log \mathbb{P}_\theta \right\rangle.$$

$$\text{and } \phi_t(\theta) = \frac{1}{t}\sum_{k=1}^t A^{t-k} \log \mathbb{p}_\theta(X_k)$$

By Lemma 34 [Distributed law of large numbers],

$$\phi_t(\theta) \overset{\mathbb{P}_0}{\to} \frac{1}{m}\sum_{i=1}^m \mathbb{P}_0 \log \mathbb{P}_\theta^i = f(\theta)\mathbf{1}.$$

which implies the coordinate-wise convergence $f_t^j(\theta) \overset{\mathbb{P}_0}{\to} f(\theta)$ for every $j \in [m]$. ■

Proof [Theorem 3] Let $\epsilon > 0$ be given. If we define $\mu_t^j(B) = \int_B \exp(-tf_t^j(\theta))\Pi(d\theta)$, then $P_t^j(B) = \frac{\mu_t^j(B)}{\mu_t^j(\Theta)}$ and $\mu_t^j(\Theta) = z_t^j < \infty$.

In the proof, we consider the $\Theta = \Theta \backslash U_{null}$ where $\pi(\theta) = 0$ for every $\theta \in U_{null}$. This null set should be of little importance because the set has zero weight under P_t^j.

We now show that there exists $\alpha > 0$ such that $\inf_{\theta \in \mathcal{U}_\epsilon^c} f_t^j(\theta) \geq f(\theta_0) + \alpha$ for all t large enough. By lemma 2, we get that there exists T such that $f_t^j(\theta) - f(\theta) < \frac{\epsilon}{2}$ for $t > T$. Given that $\inf_{\theta \in \mathcal{U}_\epsilon^c} f_t^j(\theta) - f(\theta_0) = \inf_{\theta \in \mathcal{U}_\epsilon^c}[f_t^j(\theta) - f(\theta)] + [f(\theta) - f(\theta_0)]$, $\inf_{\theta \in \mathcal{U}_\epsilon^c} f_t^j(\theta) - f(\theta_0) > \frac{\epsilon}{2}$ for $t > T$ because for $\theta \in \mathcal{U}_\epsilon^c$, $f(\theta) - f(\theta_0) > \epsilon$.

The result shows that for $t > T$, for all $\theta \in \mathcal{U}_\epsilon^c$, $exp(-t(f_t^j(\theta) - f(\theta_0) - \alpha)) \leq 1$. Therefore, for $t > T$,

$$\exp(t(f(\theta_0) + \alpha))\mu_t^j(\mathcal{U}_\epsilon^c) = \int_{\mathcal{U}_\epsilon^c} \exp(-t(f_t^j(\theta) - f(\theta_0) - \alpha))\Pi(d\theta) \leq \int_{\mathcal{U}_\epsilon^c} \Pi(d\theta) \leq 1$$

For any $\theta \in A_{\alpha/2}$, $\lim_{t\to\infty} f_t^j(\theta) - f(\theta_0) - \alpha = f(\theta) - f(\theta_0) - \alpha < -\alpha/2 < 0$. Thus, $exp(-t(f_t^j(\theta) - f(\theta_0) - \alpha)) \to \infty$ as $t \to \infty$. By Fatou's lemma, since $\Pi(A_{\alpha/2}) > 0$ by assumption 2,

$$\liminf_t e^{t(f(\theta_0)+\alpha)}\mu_t^j(A_{\alpha/2}) = \liminf_t \int_{A_{\alpha/2}} \exp(-t(f_t^j(\theta) - f(\theta_0) - \alpha))\Pi(d\theta) = \infty.$$

Since $\mu_t^j(\Theta) \geq \mu_t^j(A_{\alpha/2})$, we obtain $e^{t(f(\theta_0)+\epsilon)}\mu_t^j(\Theta) \to \infty$.

Combining the two results,

$$1 - P_t^j(\mathcal{U}_\epsilon) = P_t^j(\mathcal{U}_\epsilon^c) = \frac{\mu_t^j(\mathcal{U}_\epsilon^c)}{\mu_t^j(\Theta)} = \frac{\exp(t(f(\theta_0) + \alpha))\mu_t^j(\mathcal{U}_\epsilon^c)}{\exp(t(f(\theta_0) + \alpha))\mu_t^j(\Theta)} \to 0,$$

as $t \to \infty$. ■

Proof [Lemma 5] Let j and θ be fixed. If $\mathbb{P}_0 \ll \mathbb{P}_\theta^j$, then $D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^j) < \infty$. This implies that $\mathbb{P}_0 \log \mathbb{p}_\theta^j > -\infty$. Likewise, the expectation of $\log \mathbb{P}_0$ under $\mathbb{P}_0$ is also finite, i.e., $\mathbb{P}_0 \log \mathbb{P}_0 < \infty$.

Since the KL divergence is non-negative, $D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^j) \geq 0$, it follows that the expectation of $\log \mathbb{p}_\theta^j$ under $\mathbb{P}_0$ is less than or equal to the expectation of $\log \mathbb{P}_0$ under $\mathbb{P}_0$. Therefore, $\mathbb{P}_0 \log \mathbb{p}_\theta^j \leq \mathbb{P}_0 \log \mathbb{p}_0 < \infty$, and this completes the proof. ■

B.2 Proofs of Results in Section 5

Proof [Lemma 7] Suppose $\mathcal{U}$ is a neighborhood of θ_0, within which $\theta \mapsto \log \mathbb{P}_\theta^i$ is convex for all $i \in [m]$. Being a linear combination of convex mappings, $f_t^j(\theta)$ retains convexity in $\mathcal{U}$. As a convex function, f_t^j is Lebesgue - almost surely differentiable on $\mathcal{U}$. According to

Lemma 2, $f_t^j \xrightarrow{\mathbb{P}_0} f$, and since f is also convex, it remains differentiable a.s. on $\mathcal{U}$. Given that $\nabla f_t^j(\theta)$ is non-decreasing a.s. on $\mathcal{U}$ for every coordinate, $\nabla f_t^j(\theta_0 - \epsilon) < -\eta$ and $\hat{\theta}_t^j \leq \theta_0 - \epsilon$ implies $\nabla f_t^j(\hat{\theta}_t^j) < -\eta$, which has probability tending to 0 for every $\eta > 0$ if $f_t^j(\hat{\theta}_t^j) = o_d(1)$. This shows that for every $\epsilon, \eta > 0$,

$$P(\|\hat{\theta}_t^j - \theta_0\| > \epsilon) + o(1) \geq P(\nabla f_t^j(\theta_0 - \epsilon) < -\eta, \nabla f_t^j(\theta_0 + \epsilon) > \eta)$$

Since $\nabla f(\theta_0 - \epsilon) < 0 < \nabla f(\theta_0 + \epsilon)$, taking 2η to be the smallest coordinates of $\nabla f(\theta_0 - \epsilon)$ and $\nabla f(\theta_0 + \epsilon)$ makes the right hand side converge to 1.

■

Proof [Lemma 8] Using the consistency of $\hat{\theta}_t^j$ along with Lemma 13, given the scaling factor $\epsilon_t^j = \frac{1}{\sqrt{t}}$, we have

$$f_t^j(\hat{\theta}_t^j) = f_t^j(\theta_0) + \frac{1}{\sqrt{t}}(\hat{\theta}_t^j - \theta_0)^T V_{\theta_0} \Delta_{t,\theta}^j + \frac{1}{2}(\hat{\theta}_t^j - \theta_0)^T V_{\theta_0}(\hat{\theta}_t^j - \theta_0) + o_d(1) = f_t^j(\theta_0) + o_d(1),$$

by applications of Slusky's theorem.

By Lemma 2, $f_t^j(\theta_0)$ converges to $f(\theta_0)$ in probability so $f_t^j(\hat{\theta}_t^j) = f(\theta_0) + o_d(1)$.

Assumption 4(a) assures the existence of $\delta > 0$ such that:

$$\begin{aligned}\inf_{\theta \in B_\epsilon(\hat{\theta}_t^j)^c} (f_t^j(\theta) - f_t^j(\hat{\theta}_t^j)) &= \inf_{\theta \in B_{2\epsilon}(\theta_0)^c} (f_t^j(\theta) - f_t^j(\hat{\theta}_t^j)) \\ &= \inf_{\theta \in B_{2\epsilon}(\theta_0)^c} (f_t^j(\theta) - f(\theta_0)) + o_d(1) \\ &> 2\delta + o_d(1) \overset{t \text{ large}}{>} \delta.\end{aligned}$$

hence $\inf_{\theta \in B_\epsilon(\hat{\theta}_t^j)^c} (f_t^j(\theta) - f_t^j(\hat{\theta}_t^j)) > \delta$ with probability tending to 1 under $[\mathbb{P}_0]$. ■

Proof [Lemma 9] For any $\epsilon > 0$ and $\delta > 0$,

$$Q_t(B_\delta(0)) = \Pi_t(B_{\delta/a_t}(\theta_t)) \leq \Pi_t(B_\epsilon(\theta_0)),$$

for all t sufficiently large. Therefore, since $Q_t \to Q$ in total variational distance, we obtain

$$Q(B_\delta(0)) = \lim_{t \to \infty} Q_t(B_\delta(0)) \leq \liminf_t \Pi_t(B_\epsilon(\theta_0)).$$

Take δ arbitrarily large shows that $\lim_{t \to \infty} \Pi_t(B_\epsilon(\theta_0)) = 1$. ■

Proof [Theorem 10] For each $j \in [m]$, the sequence of estimators $\hat{\theta}_t^j$ are are chosen to satisfy the equation

$$\sum_{k=1}^{t} \sum_{i=1}^{m} [A_{ij}^{t-k}] \nabla \log p_{\hat{\theta}_t^j}^i(X_k^i) = 0.$$

The function $f_t^j(\theta)$ admits the following quadratic expansion.

$$f_t^j(\theta) = -\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]\log \mathbb{p}_\theta^i(X_k^i)$$

$$= -\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]\{\log p_{\hat\theta_t^j}^i(X_k^i) + \nabla \log p_{\hat\theta_t^j}^i(X_k^i)(\theta - \hat\theta_t^j)$$

$$-\frac{1}{2}(\theta - \hat\theta_t^j)\nabla \log p_{\hat\theta_t^j}^i(X_k^i)\left(\nabla \log p_{\hat\theta_t^j}^i(X_k^i)\right)^T(\theta - \hat\theta_t^j)^T + O_p(|\theta - \hat\theta_t^j|^3)\}$$

$$= f_t^j(\hat\theta_t^j) + \frac{1}{2}(\theta - \hat\theta_t^j)\left\{\underbrace{\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]\nabla \log p_{\hat\theta_t^j}^i(X_k^i)\left(\nabla \log p_{\hat\theta_t^j}^i(X_k^i)\right)^T}_{\hat V_t^j}\right\}(\theta - \hat\theta_t^j)^T - \underbrace{O_p(|\theta - \hat\theta_t^j|^3)}_{r_t^j(\theta - \hat\theta_t^j)}.$$

The sequence of local estimates $\hat\theta_t^j \in \Theta$ satisfies the consistency requirement. By denoting the remainder term as $r_t^j(\theta - \hat\theta_t^j)$, it immediately holds that

$$r_t^j(h) = O_p(|h|^3).$$

By Theorem 7.2 of Van der Vaart (2000), Assumption 2(c) implies the existence of the nonsingular Fisher information matrix $V_{\theta_0}^i$ for any $i \in [m]$. The average of Fisher information matrix, defined as $V_{\theta_0} = \frac{1}{m}\sum_{i=1}^m V_{\theta_0}^i$, is also nonsingular and positive definite. It remains to show that $\hat V_t^j \xrightarrow{\mathbb{P}_0} V_{\theta_0}$.

We represent the sequence of matrices $\hat V_t^j$ as:

$$\hat V_t^j = \frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]\nabla \log p_{\hat\theta_t^j}^i(X_k^i)\left(\nabla \log p_{\hat\theta_t^j}^i(X_k^i)\right)^T.$$

Using Lipschitz continuity of the gradient (Assumption 2(d)), we can decompose $\hat V_t^j$ as:

$$\hat V_t^j = \frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]\left(\nabla \log \mathbb{p}_{\theta_0}^i(X_k^i) + s^i(X_k^i)(\hat\theta_t^j - \theta_0)\right)\left(\nabla \log \mathbb{p}_{\theta_0}^i(X_k^i) + s^i(X_k^i)(\hat\theta_t^j - \theta_0)\right)^T.$$

Applying the distributive property, we can decompose this into three terms:

$$\hat V_t^j = \frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]\nabla \log \mathbb{p}_{\theta_0}^i(X_k^i)\left(\nabla \log \mathbb{p}_{\theta_0}^i(X_k^i)\right)^T$$
$$+ (\hat\theta_t^j - \theta_0)\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]s^i(X_k^i)\nabla \log \mathbb{p}_{\theta_0}^i(X_k^i)^T$$
$$+ (\hat\theta_t^j - \theta_0)(\hat\theta_t^j - \theta_0)^T\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}][s^i(X_k^i)]^2.$$

By Lemma 34, we have

$$\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]\nabla\log\mathbb{p}_{\theta_0}^{i}(X_k^i)\left(\nabla\log\mathbb{p}_{\theta_0}^{i}(X_k^i)\right)^T \overset{\mathbb{P}_0}{\to} V_{\theta_0},$$

and

$$\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}]s^i(X_k^i)\nabla\log\mathbb{p}_{\theta_0}^{i}(X_k^i)^T \overset{\mathbb{P}_0}{\to} \mathbb{P}_0[s^i\nabla\log\mathbb{p}_{\theta_0}^{i}],$$

and By the Cauchy-Schwarz inequality, we have:

$$\mathbb{P}_0[s^i\nabla\log\mathbb{p}_{\theta_0}^{i}] \leq \sqrt{\mathbb{P}_0[s^i]^2}\sqrt{V_{\theta_0}^i} < \infty,$$

thus the second term is $o(1)$.

$$\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}][s^i(X_k^i)]^2 \overset{\mathbb{P}_0}{\to} \mathbb{P}_0[s^i(X_k^i)]^2.$$

Finally, since $\hat{\theta}_t^j - \theta_0 = o_p(1)$, the second and the third term are $o_p(1)$. Combining all these results, we have

$$\hat{V}_t^j \overset{\mathbb{P}_0}{\to} V_{\theta_0}.$$

Note that $q_t^j(x) = p_t^j(\hat{\theta}_t^j + x/\sqrt{t})t^{-p/2}$. Define

$$\begin{aligned} g_t^j(x) &= \exp(-t[f_t^j(\hat{\theta}_t^j + x/\sqrt{t}) - f_t^j(\hat{\theta}_t^j)])\pi(\hat{\theta}_t^j + x/\sqrt{t}) \\ &= \exp(tf_t^j(\hat{\theta}_t^j))q_t^j(x)z_t^j t^{p/2}. \end{aligned}$$

recalling that $z_t^j < \infty$ by assumption and define

$$g_0(x) = \exp(-\frac{1}{2}x^T V_{\theta_0}x)\pi(\theta_0)$$

Let $\alpha \in (0, \lambda)$, where λ is the smallest eigenvalue of V_{θ_0}. Let $\epsilon > 0$ small enough that $\epsilon < \alpha/(2c_0), \epsilon < \epsilon_0$ and $\pi(\theta) \leq 2\pi(\theta_0)$ for all $\theta \in B_{2\epsilon}(\theta_0)$. Let $\alpha = \liminf_t \inf_{\theta\in B_\epsilon(\hat{\theta}_t^j)^c}(f_t^j(\theta) - f_t^j(\hat{\theta}_t^j))$, noting that $\delta > 0$ with asymptotic probability 1 by lemma 8. Let $A_t^j = V_t^j - \alpha I$ and $A_0 = V_{\theta_0} - \alpha I$, define

$$\begin{aligned} h_t^j(x) &= \begin{cases} \exp(-\frac{1}{2}x^T A_t^j x)2\pi(\theta_0) & \text{if } |x| < \epsilon\sqrt{t} \\ \exp(-\frac{1}{2}t\delta)2\pi(\hat{\theta}_t^j + x/\sqrt{t}) & \text{if } |x| \geq \epsilon\sqrt{t} \end{cases} \\ h_0(x) &= \exp(-\frac{1}{2}x^T A_0 x)2\pi(\theta_0) \end{aligned}$$

We show that

1. $g_t^j \to g_0$ and $h_t^j \to h_0$ almost surely
2. $\int h_t^j$ converges to $\int h_0$

3. $g_t^j = |g_t^j| \leq h_t^j$ with asymptotic probability 1, and
4. $g_t^j, g_0, h_t^j, h_0 \in L^1(dx)$ with asymptotic probability 1.

By the generalized dominated convergence theorem, these results imply that $\int g_t^j \overset{\mathbb{P}_0}{\to} \int g_0$ and $\int |(g_t^j)1/m - g_0|dx \overset{\mathbb{P}_0}{\to} 0$ with $[\mathbb{P}_0]-$probability. Assuming these results are true, we want to show how asymptotic normality follows. Since $\int q_t^j = 1$, the definition of g_t^j implies that

$$\exp(tf_t^j(\hat{\theta}_t^j))t^{p/2}z_t^j = \int g_t^j \to \int g_0 = \pi(\theta_0)\frac{(2\pi)^{p/2}}{|mV_{\theta_0}|^{1/2}}.$$

where $|V_{\theta_0}| = |det(V_{\theta_0})|$ and therefore

$$z_t^j \approx \frac{exp(-tf_t^j(\hat{\theta}_t^j))\pi(\theta_0)}{|mV_{\theta_0}|^{1/2}}(\frac{2\pi}{t})^{p/2}.$$

as $t \to \infty$. For any $a_t^j \to a \in \mathbb{R}$, we have that $\int |a_t^j g_t^j - ag_0| \to 0$ since

$$\int |a_t^j g_t^j - ag_0| \leq \int |a_t^j g_t^j - a_t^j g_0| + \int |a_t^j g_0 - ag_0| = |a_t^j| \int |g_t^j - g_0| + |a_t^j - a| \int |g_0| \to 0$$

Therefore, if we let $a_t^j = (exp(tf_t^j(\hat{\theta}_t^j))t^{p/2}z_t^j)^{-1}$ and $a = (\pi(\theta_0)\frac{(2\pi)^{p/2}}{|V_{\theta_0}|^{1/2}})^{-1}$, we have shown that $a_t^j \to a$ and thus

$$\int |q_t^j(x) - \frac{|V_{\theta_0}|^{1/2}}{(2\pi)^{p/2}}\exp(-\frac{1}{2}x^T V_{\theta_0} x)|dx \overset{\mathbb{P}_0}{\to} 0.$$

This shows asymptotic normality. Finally, the posterior consistency at θ_0 follows from Lemma 9. It remains to show statements 1 to 4 above.

1) Fix $x \in \mathbb{R}^p$. For t sufficiently large, $|x| \leq \epsilon\sqrt{t}$. It follows that

$$h_t^j(x) = \exp(-\frac{1}{2}x^T A_t^j x)2\pi(\theta) \to \exp(-\frac{1}{2}x^T V_{\theta_0} x)2\pi(\theta_0) = h_0(x).$$

since $A_t^j \to A_0$ almost surely. For g_t^j, first note that $\pi(\hat{\theta}_t^j + x/\sqrt{t}) \to \pi(\theta_0)$ since π is continuous at θ_0 and $\theta_t^j \to \theta_0$ and $x/\sqrt{t} \to 0$. By the uniform convergence of f_t^j to f_t,

$$t(f_t^j(\hat{\theta}_t^j + x/\sqrt{t}) - f_t^j(\hat{\theta}_t^j)) = -\frac{1}{2}x^T V_t^j x + tr_t^j(x/\sqrt{t}) \to -\frac{1}{2}x^T V_{\theta_0} x.$$

because $V_t^j \to V_{\theta_0}$ and $|x/\sqrt{t}| < \epsilon_0$ for all t sufficiently large. We then infer a bound on r_t^j,

$$|tr_t^j(x/\sqrt{t})| \leq tc_0|x/\sqrt{t}|^3 = c_0|x|^3/\sqrt{t} \to 0.$$

as $t \to \infty$. Therefore, $g_t^j(x) \to g_0(x)$.

2) By the definition of h_t^j, letting $B_t = B_{\epsilon\sqrt{t}}(0)$,

$$\int h_t^j = \int_{B_t} \exp(-\frac{1}{2}x^T A_t^j x)2\pi(\theta_0)dx + \int_{B_t^c} exp(-\frac{1}{2}n\delta)2\pi(\hat{\theta}_t^j + x/\sqrt{t})dx.$$

Since $A_t^j \to A_0$ almost surely and A_0 is positive definite, for all t sufficiently large, A_t^j is also positive definite and the first term equals

$$2\pi(\theta_0)\frac{(2\pi)^{p/2}}{|A_t^j|^{1/2}}P(|(A_t^j)^{-1/2}Z| < \epsilon\sqrt{t}) \overset{a.s.}{\to} 2\pi(\theta_0)\frac{(2\pi)^{p/2}}{|A_0|^{1/2}} = \int h_0.$$

where $Z \sim N(0,1)$. The second integral converges to 0, since it is nonnegative and has an upper bound as

$$\int_{\mathbb{R}^p} \exp(-\frac{1}{2}t\delta)2\pi(\hat{\theta}_t^j + x/\sqrt{t})dx = \exp^{-\frac{1}{2}n\delta} n^{p/2} \to 0.$$

using the fact that $\pi(\hat{\theta}_t^j + x/\sqrt{t})t^{-p/2}$ is the density of $\sqrt{t}(\theta - \hat{\theta}_t^j)$ where $\theta \sim \pi$.

3) With $\mathbb{P}_0-$probability going to 1, $|\theta_t^j - \theta_0| < \epsilon$ thus the bound on r_t^j applies, $\inf_{\theta \in B_\epsilon(\hat{\theta}_t^j)^c}(f_t^j(\theta) - f_t^j(\hat{\theta}_t^j)) > \delta$. Let $x \in \mathbb{R}^p$ be given again. If $|x| > \epsilon\sqrt{t}$, then $f_t^j(\hat{\theta}_t^j + x/\sqrt{t}) - f_t^j(\hat{\theta}_t^j) > \delta/2$, and thus,

$$g_t^j(x) \leq \exp(-\frac{1}{2}t\delta)\pi(\hat{\theta}_t^j + x/\sqrt{t}) = h_t^j(x).$$

In the meantime, if $|x| < \epsilon\sqrt{t}$, then $\pi(\hat{\theta}_t^j + x/\sqrt{t}) \leq 2\pi(\theta_0)$ by our choice of ϵ, and

$$\begin{aligned} t(f_t^j(\hat{\theta}_t^j + x/\sqrt{t}) - f_t^j(\hat{\theta}_t^j)) &\geq \frac{1}{2}x^T V_t^j x + tr_t^j(x/\sqrt{t}) \\ &\geq \frac{1}{2}x^T V_t^j x - \frac{1}{2}\alpha x T x = \frac{1}{2}x^T A_t^j x. \end{aligned}$$

since $|tr_t^j(x/\sqrt{t})| \leq 1/2\alpha|x|^2$, by the fact that $|x/\sqrt{t}| < \epsilon < \epsilon_0$ and $\epsilon < \alpha/(2c_0)$. Therefore,

$$\lim_{t\to\infty} \mathbb{P}_0\left(g_t^j(x) \leq h_t^j(x)\right) = 1.$$

4) Since V_{θ_0} and A_0 are positive definite, $\int g_0$ and $\int h_0$ are finite. Since $\int h_t^j \to \int h_0$, we have $\lim_{t\to\infty} \mathbb{P}_0(\int g_t^j \leq \int h_t^j < \infty) = 1$. The measurability of g_t^j, h_t^j with respect to $\mathbb{P}_0$ follows from the measurability of f_t^j and π. ■

Proof [Corollary 11] In the proof of Theorem 10, the differential in quadratic means assumption (Assumption 2(c)) is only used to establish Equation (5.2). Assumption 2(e) states that each $\log \mathbb{P}_\theta^i$ is twice continuously differentiable with nonsingular Hessian matrices. Since $f_t^j(\theta)$ is a positive linear combination of $\log \mathbb{P}_\theta^i$, $f_t^j(\theta)$ is also twice continuously differentiable with a nonsingular Hessian matrix. Therefore, Equation (5.2) holds when we replace Assumption 2(c) with Assumption 2(e). ■

Proof [Corollary 12] Denote $\langle \cdot, \cdot \rangle$ as the tensor inner product, and $\otimes$ as the tensor product. Let Assumption 2(f) hold.

We have a third - order Taylor expansion of $f_t^j(\theta)$ around $\hat{\theta}_t^j$ with nonsingular quadratic terms:

$$f_t^j(\theta) = f_t^j(\hat{\theta}_t^j) + \frac{1}{2}(\theta - \theta_t^j)^T \nabla_\theta^2 f_t^j(\hat{\theta}_t^j)(\theta - \theta_t^j) + \frac{1}{6}\langle \nabla_\theta^3 f_t^j(\tilde{\theta}), (\theta - \theta_t^j) \otimes (\theta - \theta_t^j) \otimes (\theta - \theta_t^j) \rangle,$$

for some $\tilde{\theta}$ between θ_0 and $\hat{\theta}_t^j$.

Under Assumption 2(f), we have $|\nabla_\theta^3 \log \mathbb{p}_\theta^j(x)| \le \phi^j(x)$ for integrable ϕ in a neighborhood $\mathcal{N}^j$ of θ_0. Define $\mathcal{N} = \cap_{j=1}^m \mathcal{N}^j$. The consistency assumption implies that $\mathbb{P}_0(\hat{\theta}_t^j \in \mathcal{N}) \to 1$. For all $\theta \in \mathcal{N}$, we have

$$\mathbb{P}_0|\nabla_\theta^3 f_t^j(\tilde{\theta})| \le \mathbb{P}_0 \frac{1}{t}\sum_{k=1}^t \sum_{i=1}^m [A_{ji}^{t-k}]|\nabla^3 \log \mathbb{p}_\theta^j(X_k^j)| \le \mathbb{P}_0 \frac{1}{t}\sum_{k=1}^t \sum_{i=1}^m [A_{ji}^{t-k}]\phi^j(X_k^j) < \infty.$$

By the uniform large of large number (Theorem 1.3.3, Ghosh and Ramamoorthi (2003)) and Lemma 34, we have

$$\sup_{\theta \in \mathcal{N}} \left\| \nabla_\theta^3 f_t^j(\tilde{\theta}) - \frac{1}{m}\sum_{i=1}^m \mathbb{P}_0 \nabla^3 \log \mathbb{p}_\theta^i \right\| \to 0.$$

In the limit, we have $\frac{1}{m}\sum_{i=1}^m \mathbb{P}_0|\nabla^3 \log \mathbb{p}_\theta^i| \le \frac{1}{m}\sum_{i=1}^m \mathbb{P}_0 \phi^i$, hence $\nabla_\theta^3 f_t^j(\tilde{\theta})$ is asymptotically tight and the cubic term is $o_p(1)$.

We also have that

$$\nabla_\theta^2 f_t^j(\hat{\theta}_t^j) \le \nabla_\theta^2 f_t^j(\theta_0) + \frac{1}{t}\sum_{k=1}^t \sum_{i=1}^m [A_{ji}^{t-k}]|\langle \phi^j(X_k^j), \hat{\theta}_t^j - \theta_0 \rangle|,$$

hence $\nabla_\theta^2 f_t^j(\hat{\theta}_t^j) = \nabla_\theta^2 f_t^j(\theta_0) + o_p(1)$.

Finally, we have $f_t^j(\hat{\theta}_t^j) = f_t^j(\theta_0) + o_p(1)$ by continuous mapping theorem (Theorem 18.11, Van der Vaart (2000)).

Putting the terms together, we conclude that

$$f_t^j(\theta) = f_t^j(\theta_0) + \frac{1}{2}(\theta - \theta_t^j)^T \nabla_\theta^2 f_t^j(\theta_0)(\theta - \theta_t^j) + o_p(1).$$

Thus, Equation (5.2) holds we substitute Assumption 2(c) and 2(d). ■

Proof [Lemma 13] By Theorem 7.2 from Van der Vaart (2000), Differentiability in quadratic means implies that $\mathbb{P}_0 \nabla \log \mathbb{p}_{\theta_0}^j = 0$ and the Fisher information matrix $V_{\theta_0}^j = \mathbb{P}_0 \nabla \log \mathbb{p}_{\theta_0}^j (\nabla \log \mathbb{p}_{\theta_0}^j)^T$ exists and is non-singular.

By the central limit theorem, we have

$$\frac{1}{\sqrt{t}} \sum_{k=1}^t \nabla \log \mathbb{p}_{\theta_0}^j(X_k^j) \xrightarrow{d} N(0, V_{\theta_0}^j)$$

Let's define V_{θ_0} as the average Fisher information matrix, i.e., $V_{\theta_0} = \frac{1}{m}\sum_{j=1}^m V_{\theta_0}^j$.

If we choose $\epsilon_t^j = \frac{1}{\sqrt{t}}$, then by Taylor expansion,

$$
\begin{aligned}
& t f_t^j(\theta_0 + \epsilon_t^j h) \\
& = -\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}] \log p^i_{\theta_0+\epsilon_t^j h}(X_k^i) \\
& = -\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ij}^{t-k}] \left\{ \log \mathbb{p}^i_{\theta_0}(X_k^i) + h^T \frac{\nabla \log \mathbb{p}^i_{\theta_0}(X_k^i)}{\sqrt{t}} - \frac{1}{2} h^T \frac{\nabla \log \mathbb{p}^i_{\theta_0}(X_k^i) \nabla \log \mathbb{p}^i_{\theta_0}(X_k^i)^T}{t} h + O_p(|\frac{h}{\sqrt{t}}|^3) \right\} \\
& = t f_t^j(\theta_0) - h^T \frac{\sum_{k=1}^t \sum_{i=1}^m [A_{ij}^{t-k}] \nabla \log \mathbb{p}^i_{\theta_0}(X_k^i)}{\sqrt{t}} + \frac{1}{2} h^T \frac{\sum_{k=1}^t \sum_{i=1}^m [A_{ij}^{t-k}] \nabla \log \mathbb{p}^i_{\theta_0}(X_k^i) \nabla \log \mathbb{p}^i_{\theta_0}(X_k^i)^T}{t} h \\
& - O_p(\frac{|h|^3}{\sqrt{t}}).
\end{aligned}
$$

Define Δ_{t,θ_0}^j and V_t as:

$$
\begin{aligned}
\Delta_{t,\theta_0}^j &= \frac{1}{\sqrt{t}} V_{\theta_0}^{-1} \left(\sum_{k=1}^{t}\sum_{i=1}^{m} [A_{ij}^{t-k}] \nabla \log \mathbb{p}^i_{\theta_0}(X_k^i) \right), \quad \text{and} \\
V_t^j &= \frac{1}{t} \left(\sum_{k=1}^{t}\sum_{i=1}^{m} [A_{ij}^{t-k}] \nabla \log \mathbb{p}^i_{\theta_0}(X_k^i) \nabla \log \mathbb{p}^i_{\theta_0}(X_k^i)^T \right).
\end{aligned}
$$

Substituting these terms in the Taylor expansion, we have:

$$
t f_t^j(\theta_0 + \epsilon_t^j h) = t f_t^j(\theta_0) - h^T V_{\theta_0} \Delta_{t,\theta_0}^j + \frac{1}{2} h^T V_t^j h + o_p(|h^2|).
$$

By Lemma 35 (distributed central limit theorem), we have

$$
\Delta_{t,\theta_0}^j \xrightarrow{d} N(0, (m V_{\theta_0})^{-1}),
$$

thus Δ_{t,θ_0}^j is bounded in $[\mathbb{P}_0]$-probability by Prokhorov's theorem (Theorem 2.4, Van der Vaart (2000)).

By Lemma 34 (distributed law of large number),

$$
V_t^j \xrightarrow{\mathbb{P}_0} V_{\theta_0}.
$$

This establishes the stochastic LAN:

$$
\sup_{h \in K} \left| -t f_t^j(\theta_0 + \epsilon_t^j h) + t f_t^j(\theta_0) - h^T V_{\theta_0} \Delta_{t,\theta_0}^j + \frac{1}{2} h^T V_{\theta_0} h \right| = \sup_{h \in K} o_p(|h|) \to 0.
$$

■

Proof [Theorem 14] Since $P_t^j(A)$ is uniformly bounded by 1 in L_∞, the set $\{P_t^j(A)\}_{t \in \mathbb{N}}$ is uniformly integrable. Then assumption (5.6) implies that

$$
\mathbb{P}_0 P_t^j \left(\theta : d(\theta, \theta_0) \geq M_t \epsilon_t^j \right) \to 0.
$$

The rest of the proof is split into two parts: in the first part, we prove the assertion conditional on an arbitrary compact set $K \subset \mathbb{R}^p$, and in the second part, we use this to prove the asymptotic normality convergence. Throughout the proof, we denote the posterior for $h_t^j = (\theta - \theta_0)/\epsilon_t^j$ when $\theta \sim P_t^j$ by Q_t^j. For all Borel sets B, the posterior for h_t^j follows from that for θ by

$$Q_t^j(h_t^j \in B) = P_t^j\left(\theta - \theta_0 \in \epsilon_t^j B\right).$$

The assumption that (5.5) holds at θ_0 means that for all $j \in [m]$, there exists non-singular $V_{\theta_0}^j$ and asymptotically tight sequence Δ_{t,θ_0}^j such that for every compact set $K \subset \mathbb{R}^p$, as $t \to \infty$,

$$\sup_{h \in K}\left|-tf_t^j(\theta_0 + \epsilon_t^j h) + tf_t^j(\theta_0) - h^T V_{\theta_0}^j \Delta_{t,\theta_0}^j + \frac{1}{2}h^T V_{\theta_0}^j h\right| \xrightarrow{\mathbb{P}_0} 0. \tag{B.1}$$

We denote the normal distribution $N(\Delta_{t,\theta}^j, V_{\theta_0}^{-1})$ by Φ_t^j. For a compact subset $K \subset \mathbb{R}^p$ such that $Q_t^j(h_t^j \in K) > 0$, we define a conditional measure $Q_{K,t}^j$ of Q_t^j by $Q_{K,t}^j(B) = Q_t^j(B \cap K)/Q_t^j(K)$. Similarly, we define a conditional measure $\Phi_{K,t}^j$ corresponding to Φ_t^j.

Let $K \subset \mathbb{R}^p$ be a compact subset of $\mathbb{R}^p$. For every open neighborhood $U \subset \Theta$ of θ_0, $\theta_0 + K\epsilon_t^j \subset U$ for large enough t. Since θ_0 is an interior point of Θ, for large enough t, the random function $\psi_t^j : K \times K \mapsto \mathbb{R}$

$$\psi_t^j(h,h') = \left|1 - \frac{\phi_t(h')}{\phi_t(h)}\frac{q_t^j(h')}{q_t^j(h)}\exp\left(tf_t^j(\theta_0 + \epsilon_t^j h) - tf_t^j(\theta_0 + \epsilon_t^j h')\right)\right|_+$$

is well-defined, where $\phi_t : K \to \mathbb{R}$ is the Lebesgue density of the distribution $N(\Delta_{t,\theta_0}^j, V_{\theta_0}^{-1})$, $q_t^j : K \to \mathbb{R}$ is the Lebesgue density of the prior for the centered and rescaled parameter h_t^j, and $f_t^j : K \to \mathbb{R}$ is the random functions defined in (2.10).

By the distributed LAN assumption (B.1), we have for every $h \in K$,

$$tf_t^j(\theta_0 + \epsilon_t^j h) - tf_t^j(\theta_0) = h^T V_{\theta_0}\Delta_{t,\theta_0}^j - \frac{1}{2}h^T V_{\theta_0} h + o_{\mathbb{P}_0}(1).$$

For any two sequences $h_t, h_t' \in K$, $\lim_{t\to\infty} q_t^j(h_t)/q_t^j(h_t') = 1$ implies that

$$\begin{aligned}
&\log\left(\frac{\phi_t(h_t')}{\phi_t(h_t)}\frac{q_t^j(h_t)}{q_t^j(h_t')}\right) + tf_t^j(\theta_0 + \epsilon_t^j h_t) - tf_t^j(\theta_0 + \epsilon_t^j h_t')\\
&= (h_t' - h_t)^T V_{\theta_0}\Delta_{t,\theta_0}^j - \frac{1}{2}h_t^T V_{\theta_0} h_t + \frac{1}{2}{h_t'}^T V_{\theta_0} h_t' + o_{\mathbb{P}_0}(1)\\
&+ \frac{1}{2}(h_t' - \Delta_{t,\theta_0}^j)^T V_{\theta_0}(h_t' - \Delta_{t,\theta_0}^j) - \frac{1}{2}(h_t - \Delta_{t,\theta_0}^j)^T V_{\theta_0}(h_t - \Delta_{t,\theta_0}^j)\\
&= o_{\mathbb{P}_0}(1),
\end{aligned}$$

as $t \to \infty$.

Since all functions ψ_t^j depend continuously on (h,h') and $K \times K$ is compact, we conclude that,

$$\sup_{h,h' \in K} \psi_t^j(h,h') \xrightarrow{\mathbb{P}_0} 0, \text{ as } t \to \infty, \tag{B.2}$$

where the convergence is in outer probability.

Assume that K contains a neighborhood of 0 (to guarantee that $\Phi_t^j(K) > 0$) and let Ξ_t^j denote the event that $Q_t^j(K) > 0$. Let $\eta > 0$ be given and define the events:

$$\Omega_t^j = \left\{ \sup_{h,h' \in K} \psi_t^j(h,h') \leq \eta \right\}.$$

Consider the inequality (recall that the total-variation norm $|| \cdot ||_{TV}$ is bounded by 2):

$$\mathbb{P}_0 \| Q_{K,t}^j - \Phi_{K,t}^j \|_{TV} 1_{\Xi_t^j} \leq \mathbb{P}_0 \| Q_{K,t}^j - \Phi_{K,t}^j \|_{TV} 1_{\Omega_t^j \cap \Xi_t^j} + 2\mathbb{P}_0(\Xi_t^j \setminus \Omega_t^j). \tag{B.3}$$

As a result of Equation (B.2), the second term is $o(1)$ as $t \to \infty$. The first term on the r.h.s. is calculated as follows:

$$\begin{aligned}\frac{1}{2}\mathbb{P}_0 \| Q_{K,t}^j - \Phi_{K,t}^j \|_{TV} 1_{\Omega_t^j \cap \Xi_t^j} &= \mathbb{P}_0 \int \left(1 - \frac{d\Phi_{K,t}^j}{dQ_{K,t}^j}\right)_+ dQ_{K,t}^j 1_{\Omega_t^j \cap \Xi_t^j} \\ &= \mathbb{P}_0 \int_K \left(1 - \int_K \frac{\phi_{K,t}^j(h) q_t^j(h') s_t^j(h')}{\phi_{K,t}^j(h') q_t^j(h) s_t^j(h)} d\Phi_{K,t}^j(h')\right)_+ dQ_{K,t}^j(h) 1_{\Omega_t^j \cap \Xi_t^j},\end{aligned}$$

where $s_t^j : K \to \mathbb{R}$ is the function defined as $s_t^j(h) = \exp(t f_t^j(\theta_0 + \epsilon_t^j h))$, and $q_t^j : K \to \mathbb{R}$ is the Lebesgue density of the prior for the centered and rescaled parameter h_t^j.

Note that for all $h, h' \in K$, $\phi_{K,t}^j(h)/\phi_{K,t}^j(h') = \phi_t(h)/\phi_t(h')$, since on K, $\phi_{K,t}^j$ differs from ϕ_t only by a normalization factor. Using Jensen's inequality (with respect to the $\Phi_{K,t}^j$-expectation) for the convex function $x \mapsto (1-x)_+$, we derive:

$$\begin{aligned}&\frac{1}{2}\mathbb{P}_0 \| Q_{K,t}^j - \Phi_{K,t}^j \|_{TV} 1_{\Omega_t^j \cap \Xi_t^j} \\ &\leq \mathbb{P}_0 \int \left(1 - \frac{s_t^j(h') q_t^j(h') \phi_t(h')}{s_t^j(h') q_t^j(h') \phi_t(h')}\right)_+ d\Phi_{K,t}^j(h') dQ_{K,t}^j(h') 1_{\Omega_t^j \cap \Xi_t^j} \\ &\leq \mathbb{P}_0 \int_{\Omega_t^j \cap \Xi_t^j} \sup_{h,h' \in K} \psi_t^j(h,h') d\Phi_{K,t}^j(h') dQ_{K,t}^j(h') \\ &\leq \eta.\end{aligned}$$

Combination with (B.3) shows that for all compact $K \subset \mathbb{R}^p$ containing a neighborhood of 0,

$$\mathbb{P}_0 \| Q_{K,t}^j - \Phi_{K,t}^j \|_{TV} 1_{\Xi_t^j} \to 0. \tag{B.4}$$

Now, let (K_n) be a sequence of balls centered at 0 with radii $M_n \to \infty$. For each $n \geq 1$, the Equation (B.4) holds. Hence, the intuition is that if we choose a sequence of balls (K_t) that traverses the sequence K_n slowly enough, the convergence of the expected total variational distance to zero can still be guaranteed. Moreover, the corresponding events $\Xi_t^j = \{Q_t^j(K_t) > 0\}$ satisfy $\mathbb{P}_0(\Xi_t^j) \to 1$ as a result of assumption (5.6).

We conclude that there exists a sequence of radii (M_t) such that $M_t \to \infty$ and $\mathbb{P}_0 \| Q_{K_t,t}^j - \Phi_{K_t,t}^j \|_{TV} \overset{\mathbb{P}_0}{\to} 0$ (where the conditional probabilities on the l.h.s. are well-defined on sets of probability growing to one). The total variation distance between a measure and its conditional version on a set K satisfies $\|Q - Q_K\|_{TV} \leq 2Q(K^c)$. Combining this with assumption (5.6) and the Hellinger integral test, we conclude that

$$\mathbb{P}_0 \|Q_t^j - \Phi_t^j\|_{TV} \to 0,$$

which implies the main result. ■

Proofs of Results in Section 6

Lemma 30 *For any function $f \geq 0$ and two probability measures P, Q, we have*

$$\int f(x) dQ(x) \leq D_{KL}(Q \parallel P) + \log \int \exp(f(x)) dP(x).$$

Proof [Lemma 30] By the definition of KL divergence,

$$\begin{aligned} D_{KL}(Q \parallel P) + \log \exp(f(x)) dP(x) &= \int \log \left(\frac{dQ(x) \int \exp(f(y)) dP(y)}{dP(x)} dQ(x) \right) \\ &= \int \log \left(\frac{dQ(x) \int \exp(f(y)) dP(y)}{\exp(f(x)) dP(x)} dQ(x) \right) + \int f(x) dQ(x) \\ &= D(Q \parallel P') + \int f(x) dQ(x) \geq \int f(x) dQ(x) \end{aligned}$$

where P' is given by

$$dP'(x) = \frac{\exp(f(x)) dP(x)}{\int \exp(f(y) dP(y)}.$$

■

Lemma 31 *For P_t^j defined in (2.9) and P_t defined in (6.2), we have the following upper bound on the expected loss under P_t^j and P_0.*

$$\mathbb{P}_0 P_t^j d(\theta, \theta_0) \leq \inf_{a>0} \frac{1}{amt} \left[\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) + \log \mathbb{P}_{\theta_0} P_t e^{amtd(\theta, \theta_0)} + \frac{1}{m} \sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j) \right].$$

Proof [Lemma 31] By Lemma 30, we have

$$amt P_t^j d(\theta, \theta_0) \leq D_{KL}(P_t^j \parallel P_t) + \log P_t \exp(amtd(\theta, \theta_0)),$$

for all $a > 0$. Taking expectations on both sides, we have

$$amt \mathbb{P}_0 P_t d(\theta, \theta_0) \leq \mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) + \mathbb{P}_0 \log P_t \exp(amtd(\theta, \theta_0)),$$

By Lemma 30 and the tenderization property of KL divergence, we have

$$\mathbb{P}_0 \log P_t \exp(amtd(\theta, \theta_0)) \leq \left(\frac{1}{m} \sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j) \right) + \log \mathbb{P}_{\theta_0} P_t \exp(amtd(\theta, \theta_0)).$$

Putting it all together, we have

$$\mathbb{P}_0 P_t^j d(\theta,\theta_0) \le \inf_{a>0} \frac{1}{amt} \left(\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) + \frac{1}{m}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j) + \log \mathbb{P}_{\theta_0} P_t \exp\left(amtd(\theta,\theta_0)\right) \right).$$

■

Lemma 32 *Let the assumptions of Theorem 15 hold. We have*

$$\mathbb{P}_{\theta_0} P_t(d(\theta,\theta_0) > C_1\epsilon^2) \le \exp(-C_0 t\epsilon^2) + \exp(-\lambda t\epsilon^2) + 2\exp(-t\epsilon^2),$$

for all $\epsilon \ge \epsilon_{m,t}^2$*, where* $\lambda = \rho - 1$ *in assumption* (C3).

Proof [Lemma 32] We define the sets

$$U = \{\theta : d(\theta,\theta_0) > C_1\epsilon^2\}, \qquad K_t = \{\theta : \frac{1}{m}\sum_{j=1}^m D_{1+\lambda}(\mathbb{P}_{\theta_0}^j \parallel \mathbb{P}_\theta^j) \le C_3\epsilon_{m,t}^2\}.$$

Let $\tilde{\Pi}$ be a probability measure defined as $\tilde{\Pi}(B) = \frac{\Pi(B\cap K_t)}{\Pi(K_t)}$. Define the event

$$A_t = \{X^{(mt)} : \int_\Theta \frac{\mathbb{P}_\theta}{\mathbb{P}_{\theta_0}}(X^{(mt)}) d\tilde{\Pi}(\theta) \le \exp(-(C_3+1)t\epsilon^2)\},$$

Let Π_t^j and ϕ_t be the set and testing function in (C1). We bound $\mathbb{P}_{\theta_0} P_t(U)$ by

$$\begin{aligned}
\mathbb{P}_{\theta_0} P_t(U) &\le \mathbb{P}_{\theta_0}\phi_t(X^{(mt)}) + \mathbb{P}_{\theta_0}(A_t) + \mathbb{P}_{\theta_0}(1-\phi_t(X^{(mt)}))P_t(U) I_{(A_t)^c} \\
&= \mathbb{P}_{\theta_0}\phi_t(X^{(mt)}) + \mathbb{P}_{\theta_0}(A_t) + \mathbb{P}_{\theta_0}(1-\phi_t(X^{(mt)})) I_{(A_t)^c} \frac{\int_U \frac{\mathbb{P}_\theta}{\mathbb{P}_{\theta_0}}(X^{(mt)}) d\Pi(\theta)}{\int_\Theta \frac{\mathbb{P}_\theta}{\mathbb{P}_{\theta_0}}(X^{(mt)}) d\Pi(\theta)}.
\end{aligned}$$

By (C1), we bound the first term by

$$\mathbb{P}_{\theta_0}\phi_t(X^{(mt)}) \le \exp(-C_0 t\epsilon^2).$$

By Jensen's inequality, we have

$$\begin{aligned}
D_\rho(\mathbb{P}_{\theta_0} \parallel \mathbb{P}_\theta) &= \sum_{i=1}^m \frac{1}{\rho-1} \log \int [dP_{\theta_0}^i]^{\frac{\rho}{m}} [dP_{\theta_0}^i]^{\frac{1}{m}-\frac{\rho}{m}} d\mu \\
&\le \frac{1}{m}\sum_{i=1}^m \frac{1}{\rho-1} \log \int [dP_{\theta_0}^i]^{\rho} [dP_{\theta_0}^i]^{1-\rho} d\mu = \frac{1}{m}\sum_{i=1}^m D_\rho(\mathbb{P}_{\theta_0}^i \parallel \mathbb{P}_\theta^i)
\end{aligned}$$

Using the definitions of event A_t and Markov inequality, we have

$$\begin{aligned}
\mathbb{P}_{\theta_0}(A_t) &= \mathbb{P}_{\theta_0}\left(\left(\int_\Theta \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\tilde{\Pi}(\theta)\right)^{-\lambda} < \exp(\lambda(C_3+1)t\epsilon^2\right) \\
&\leq \exp(-\lambda(C_3+1)t\epsilon^2)\mathbb{P}_{\theta_0}\left(\int_\Theta \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\tilde{\Pi}(\theta)\right)^{-\lambda} \\
&\leq \exp(-\lambda(C_3+1)t\epsilon^2)\int_\Theta \left(\int_{\mathcal{X}^{mt}} \frac{(\mathbb{p}_{\theta_0}(x^{(mt)}))^\lambda}{(\mathbb{p}_\theta(x^{(mt)}))^\lambda} d\mathbb{P}_{\theta_0}(x^{(mt)})\right) d\tilde{\Pi}(\theta) \\
&= \exp(-\lambda(C_3+1)t\epsilon^2)\int_\Theta \exp(\lambda D_{1+\lambda}(\prod_{k=1}^t \mathbb{P}_{\theta_0} \parallel \prod_{k=1}^t \mathbb{P}_\theta))d\tilde{\Pi}(\theta) \\
&= \exp(-\lambda(C_3+1)t\epsilon^2)\int_{K_t} \exp(\lambda t D_{1+\lambda}(\mathbb{P}_{\theta_0} \parallel \mathbb{P}_\theta))d\tilde{\Pi}(\theta) \\
&\leq \exp(-\lambda(C_3+1)t\epsilon^2)\int_{K_t} \exp(\lambda t \frac{1}{m}\sum_{j=1}^m D_{1+\lambda}(\mathbb{P}^j_{\theta_0} \parallel \mathbb{P}^j_\theta))d\tilde{\Pi}(\theta) \\
&\leq \exp(-\lambda(C_3+1)t\epsilon^2 + \lambda C_3 t\epsilon^2_{m,t}) \\
&\leq \exp(-\lambda t\epsilon^2).
\end{aligned}$$

Let's analyze the third term. Conditioned on $(A_t)^c$, we have

$$\int_\Theta \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\Pi(\theta) \geq \Pi(K_t)\int_\Theta \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\tilde{\Pi}(\theta) \geq \exp(-(C_2+C_3+1)t\epsilon^2),$$

where the last inequality follows from C3. It follows that

$$\begin{aligned}
&\mathbb{P}_{\theta_0}(1-\phi_t(X^{(mt)}))I_{(A_t)^c}\frac{\int_U \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\Pi(\theta)}{\int_\Theta \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\Pi(\theta)} \\
&\leq \exp((C_2+C_3+1)t\epsilon^2)\mathbb{P}_{\theta_0}\int_U \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\Pi(\theta) \\
&\leq \exp((C_2+C_3+1)t\epsilon^2)\left[\int_{U\cap\Theta_t(\epsilon)} \mathbb{P}_\theta(1-\phi_t(X^{(mt)}))d\Pi(\theta) + \Pi\left(\Theta_t(\epsilon)^c\right)\right] \\
&\leq \exp((C_2+C_3+1)t\epsilon^2)(\exp(-Ct\epsilon^2)+\exp(-Ct\epsilon^2)),
\end{aligned}$$

where the last inequality follows from C1 and C2. Since $C_0 > C_3 + C_2 + 2$, we obtain the upper bound for the third term.

$$\mathbb{P}_{\theta_0}(1-\phi_t(X^{(mt)}))I_{(A_t)^c}\frac{\int_U \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\Pi(\theta)}{\int_\Theta \frac{\mathbb{p}_\theta}{\mathbb{p}_{\theta_0}}(X^{(mt)})d\Pi(\theta)} \leq 2\exp(-t\epsilon^2).$$

Putting the bounds all together, we have

$$\mathbb{P}_{\theta_0}P_t(U) \leq \exp(-C_0 t\epsilon^2) + \exp(-\lambda t\epsilon^2) + 2\exp(-t\epsilon^2)$$

■

Lemma 33 (Sub-exponential Bound) *Let the random variable X satisfies*

$$\mathbb{P}(X \geq \delta) \leq c_1 \exp(-c_2\delta),$$

for all $\delta \geq \delta_0 > 0$. Then, for any $0 < a \leq \frac{1}{2}c_2$, we have

$$\mathbb{E}\exp(aX) \leq \exp(a\delta_0) + c_1,$$

Proof [Lemma 33] Define $Y = \exp(aX)$ for some $0 < a \leq \frac{1}{2}c_2$. For any $C > 0$,

$$\mathbb{E}Y \leq C + \int_C^\infty \mathbb{P}(Y \geq y)dy = C + \int_C^\infty \mathbb{P}(X \geq \frac{1}{a}\log y)dy \leq C + c_1 \int_C^\infty y^{-c_2/a}dy.$$

Take $C = \exp(a\delta_0)$. SInce $a \leq \frac{1}{2}c_2$, we have

$$\mathbb{E}Y \leq \exp(a_t) + c_1\exp(-a\delta_0) \leq \exp(a\delta_0) + c_1.$$

■

Proof [Proof of Theorem 15] By Lemma 32, for all $\delta \geq \delta_0$, we have

$$\mathbb{P}_{\theta_0}P_t(d(\theta,\theta_0) > \delta) \leq c_1\exp(-c_2\delta).$$

Take $c_1 = 4, c_2 = mt\min(\lambda,1)/C_1$ as $C_0 > C_1 + C_2 + 2 > 1$ and $\delta_0 = C_1 m\epsilon_{m,t}^2$. By Lemma 33, we have

$$\mathbb{P}_{\theta_0}P_t\exp(amtd(\theta,\theta_0)) \leq \exp(amtC_1\epsilon_{m,t}^2) + 4,$$

for all $a \leq \min(\lambda,1)/(2C_1)$.

Take $a = \min(\lambda,1)/(2C_1)$. By Lemma 31, we have

$$\begin{aligned}
\mathbb{P}_0 P_t^j d(\theta,\theta_0) &\leq \frac{mt\gamma_{j,m,t}^2 + \log(4+\exp(aC_1mt\epsilon_{m,t}^2)) + \frac{1}{m}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j)}{amt} \\
&\leq \frac{\gamma_{j,m,t}^2}{a} + C_1\epsilon_{m,t}^2 + \frac{4\exp(-aC_1mt\epsilon_{m,t}^2)}{amt} + \frac{\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j)}{am^2t} \\
&\leq C\left(\epsilon_{m,t}^2 + \gamma_{j,m,t}^2 + \frac{1}{m^2t}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j)\right),
\end{aligned}$$

for some C that depends only on C_0, C_1, λ. ■

Proof [Corollary 16] The first result is a consequence of Markov's inequality.

$$\begin{aligned}
&\mathbb{P}_0 P_t^j\left(d(\theta,\theta_0) > M_t(\epsilon_{m,t}^2 + \gamma_{j,m,t}^2 + \frac{1}{m^2t}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j))\right) \\
\leq\ & \frac{\mathbb{P}_0 P_t^j d(\theta,\theta_0)}{M_t(\epsilon_{m,t}^2 + \gamma_{j,m,t}^2 + \frac{1}{m^2t}\sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j))} \leq \frac{C}{M_t} \to 0
\end{aligned}$$

The second result follows from Jensen's inequality

$$\mathbb{P}_0 d(P_t^j \theta, \theta_0) \le \mathbb{P}_0 P_t d(\theta, \theta_0) \le C(\epsilon_{m,t}^2 + \gamma_{j,m,t}^2 + \frac{1}{m^2 t} \sum_{j=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j))$$

■

Proof [Lemma 17] We have

$$\begin{aligned}
\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) &= \mathbb{P}_0 P_t^j \left(\sum_{k=1}^t \sum_{i=1}^m [A_{ij}^{t-k}] \log \mathbb{P}_\theta^i(X_k^i) - \frac{1}{m} \sum_{k=1}^t \sum_{i=1}^m \log \mathbb{P}_\theta^i(X_k^i) \right) \\
&= P_t^j \mathbb{P}_0 \left(\sum_{k=1}^t \sum_{i=1}^m [A_{ij}^{t-k}] \log \mathbb{P}_\theta^i(X_k^i) - \frac{1}{m} \sum_{k=1}^t \sum_{i=1}^m \log \mathbb{P}_\theta^i(X_k^i) \right) \\
&= P_t^j \left(\sum_{k=1}^t \sum_{i=1}^m [A_{ij}^{t-k}] \mathbb{P}_0 \log \mathbb{P}_\theta^i - \frac{1}{m} \sum_{k=1}^t \sum_{i=1}^m \mathbb{P}_0 \log \mathbb{P}_\theta^i \right) \\
&= \sum_{k=1}^t \sum_{i=1}^m \left\{ [A_{ij}^{t-k}] - \frac{1}{m} \right\} P_t^j \mathbb{P}_0 \log \mathbb{P}_\theta^i \\
&= \sum_{k=1}^t \sum_{i=1}^m \left\{ [A_{ij}^{t-k}] - \frac{1}{m} \right\} \left\{ \mathbb{P}_0 \log \mathbb{p}_0 - P_t^j D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right\}.
\end{aligned}$$

For each j, we have

$$\begin{aligned}
\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) &\le \sum_{k=1}^t \sum_{i=1}^m |[A_{ij}^{t-k}] - \frac{1}{m}| \left| \mathbb{P}_0 \log \mathbb{p}_0 - \max_{i \in [m]} \inf_{\theta \in \Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right| \\
&\le \sum_{k=1}^t \sum_{i=1}^m |[A_{ij}^{t-k}] - \frac{1}{m}| \left\{ |\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i \in [m]} \inf_{\theta \in \Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right\}.
\end{aligned}$$

By Lemma 1, we get

$$\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) \le \frac{16 m^2 \log m}{\nu} \left(|\mathbb{P}_0 \log \mathbb{p}_0| + \frac{1}{m} \sum_{i=1}^m D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^i) \right).$$

By Assumption 2(a), we have

$$\gamma_{j,m,t}^2 \le \frac{16 m \log m}{\nu t} \left(|\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i \in [m]} \inf_{\theta \in \Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right).$$

■

B.3 Proofs of Results in Section 7

Proof [Proposition 19]

Define $\mathcal{T} \subseteq [t]$ as the set of τ where $G_\tau = G$. This allows us to break the left-hand side of the inequality into two parts:

$$\sum_{k=1}^{t}\sum_{j=1}^{m}\left|\left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} - \frac{\lambda}{m}\right| = \sum_{k=1}^{t}\sum_{j=1}^{m}\left|\left[\prod_{\tau\in[k,t-1]\cap\mathcal{T}} A\right]_{ij} - \frac{\lambda}{m}\right| = \sum_{k=1}^{t}\sum_{j=1}^{m}\left|\left[A^{|[k,t-1]\cap\mathcal{T}|}\right]_{ij} - \frac{\lambda}{m}\right|.$$

As $t\to\infty$, $|\mathcal{T}|/t \overset{a.s.}{\to} \lambda$ and $|[k,t-1]\cap\mathcal{T}| \overset{a.s.}{\to} \lambda(t-k)$. Then with probability 1, we have

$$\limsup_{t\to\infty}\sum_{k=1}^{t}\sum_{j=1}^{m}\left|\left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} - \frac{\lambda}{m}\right| = \limsup_{t\to\infty}\sum_{k=1}^{t}\sum_{j=1}^{m}\left|\left[A^{\lambda(t-k)}\right]_{ij} - \frac{\lambda}{m}\right|.$$

Define $\delta = \max(|\lambda_2(A)|, |\lambda_m(A)|)$. The spectral radius of A is 1, thus $\delta < 1$ by Perron-Frobenius theorem. Under the assumptions 1, by the standard property of stochastic matrices (see e.g. Rosenthal (1995)), the diagonalizable matrix A satisfies

$$\|e_i^T A^{\lambda t} - \frac{1}{m}\mathbf{1}\|_1 \le m\delta^{\lambda t}. \tag{B.5}$$

for any $i \in [m]$, using the fact that $\frac{1}{m}\mathbf{1}$ is the stationary distribution of the Markov chain with transition matrix A.

Assume that $\lambda \ge \frac{c}{m}$. For any $t - k \ge \tilde{t} = \frac{\log\frac{m}{c}+\log\lambda}{-\lambda\log\delta}$,

$$m\delta^{\lambda(t-k)} \le m\delta^{\lambda\tilde{t}} \le \frac{m}{\frac{m}{c}\lambda} \le \frac{c}{\lambda}.$$

Since $\|e_i^T A^{\lambda t} - \frac{1}{m}\mathbf{1}\|_1 \le 2$ by the double stochasticity of A, we use (B.7) to break the quantity of interest $\sum_{k=1}^{t}\sum_{j=1}^{m}|[A_{ij}^{\lambda(t-k)}] - \frac{1}{m}|$ into two parts. For any $i \in [m]$,

$$\begin{aligned}
\sum_{k=1}^{t}\sum_{j=1}^{m}|[A_{ij}^{\lambda(t-k)}] - \frac{1}{m}| &= \sum_{k=1}^{t}\|e_i^T A^{\lambda(t-k)} - \frac{1}{m}\mathbf{1}\|_1 \\
&= \sum_{k=1}^{t-\tilde{t}}\|e_i^T A^{\lambda(t-k)} - \frac{1}{m}\mathbf{1}\|_1 + \sum_{k>t-\tilde{t}}^{t}\|e_i^T A^{\lambda(t-k)} - \frac{1}{m}\mathbf{1}\|_1 \\
&\le \sum_{k=1}^{t-\tilde{t}} m\delta^{\lambda(t-k)} + 2\tilde{t} \le \frac{m\delta^{\lambda\tilde{t}}}{1-\delta} + 2\tilde{t} \le \frac{c}{\lambda(1-\delta)} + \frac{2\log\frac{m}{c} + 2\log\lambda}{-\lambda\log\delta}.
\end{aligned}$$

Noting that $1-\delta \le -\log\delta$ and $\delta = 1 - \frac{\nu}{4m^2}$ for any doubly stochastic matrix A, we get

$$\sum_{k=1}^{t}\sum_{j=1}^{m}|\left[A^{\lambda(t-k)}\right]_{ij} - \frac{1}{m}| \le \frac{c + 2\log\frac{m}{c} + \log\lambda}{\lambda(1-\delta)} \le \frac{4m^2c + 8m^2\log\frac{m}{c} + 8m^2\log\lambda}{\lambda\nu}.$$

When $\lambda \geq \frac{2}{m}$, the optimal bound is achieved when $c = \frac{1}{2}$,

$$\limsup_{t\to\infty} \sum_{k=1}^{t} \sum_{j=1}^{m} |\left[A^{\lambda(t-k)}\right]_{ij} - \frac{1}{m}| \leq \frac{8m^2(1+\log\frac{m}{2}) + 8m^2\log\lambda}{\lambda\nu} \leq \frac{16m^2\log m + 8m^2\log\lambda}{\lambda\nu}.$$

When $\lambda < \frac{2}{m}$, the optimal bound is achieved when $c = \lambda$,

$$\limsup_{t\to\infty} \sum_{k=1}^{t} \sum_{j=1}^{m} |\left[A^{\lambda(t-k)}\right]_{ij} - \frac{1}{m}| \leq \frac{4m^3}{\nu}.$$

Finally, when $\lambda = 0$,

$$\limsup_{t\to\infty} \sum_{k=1}^{t} \sum_{j=1}^{m} \left| \left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} - \frac{\lambda}{m} \right| = \limsup_{t\to\infty} \sum_{k=1}^{t} \frac{m-1}{m} = \infty.$$

■

Proof [Corollary 20] We have

$$\begin{aligned}
\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) &= \mathbb{P}_0 P_t^j \left(\sum_{k=1}^{t} \sum_{i=1}^{m} \left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} \log \mathbb{P}_\theta^i(X_k^i) - \frac{1}{m} \sum_{k=1}^{t} \sum_{i=1}^{m} \log \mathbb{P}_\theta^i(X_k^i) \right) \\
&= P_t^j \mathbb{P}_0 \left(\sum_{k=1}^{t} \sum_{i=1}^{m} \left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} \log \mathbb{P}_\theta^i(X_k^i) - \frac{1}{m} \sum_{k=1}^{t} \sum_{i=1}^{m} \log \mathbb{P}_\theta^i(X_k^i) \right) \\
&= P_t^j \left(\sum_{k=1}^{t} \sum_{i=1}^{m} \left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} \mathbb{P}_0 \log \mathbb{P}_\theta^i - \frac{1}{m} \sum_{k=1}^{t} \sum_{i=1}^{m} \mathbb{P}_0 \log \mathbb{P}_\theta^i \right) \\
&= \sum_{k=1}^{t} \sum_{i=1}^{m} \left\{ \left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} - \frac{1}{m} \right\} P_t^j \mathbb{P}_0 \log \mathbb{P}_\theta^i \\
&= \sum_{k=1}^{t} \sum_{i=1}^{m} \left\{ \left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} - \frac{1}{m} \right\} \left\{ \mathbb{P}_0 \log \mathbb{p}_0 - P_t^j D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right\}.
\end{aligned}$$

For each j, we have

$$\begin{aligned}
\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) &\leq \sum_{k=1}^{t} \sum_{i=1}^{m} |\left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} - \frac{1}{m}| \left| \mathbb{P}_0 \log \mathbb{p}_0 - \max_{i\in[m]} \inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right| \\
&\leq \sum_{k=1}^{t} \sum_{i=1}^{m} |\left[\prod_{\tau=k}^{t-1} A_\tau\right]_{ij} - \frac{1}{m}| \left\{ |\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i\in[m]} \inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right\}.
\end{aligned}$$

This implies that

$$\mathbb{P}_0 D_{KL}(P_t^j \parallel P_t) \leq \limsup_{t\to\infty} \sum_{k=1}^{t} \sum_{j=1}^{m} \left| \left[\prod_{\tau=k}^{t-1} A_\tau \right]_{ij} - \frac{1}{m} \right| \left(|\mathbb{P}_0 \log \mathbb{p}_0| + \frac{1}{m} \sum_{i=1}^{m} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^i) \right).$$

Applying Proposition 19 gives us the desired results. ■

Proof [Corollary 21] Since P_t does not depend on the communication graph, Theorem 15 still holds. This implies the following upper bound on the contraction rate:

$$\mathbb{P}_0 P_t^j d(\theta, \theta_0) \lesssim \epsilon_{m,t}^2 + \gamma_{j,m,t}^2 + \frac{1}{m^2 t} \sum_{j=1}^{m} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j). \tag{B.6}$$

The only term depending on the communication graph is $\gamma_{j,m,t}^2$. By Corollary 21, we have the following upper bounds with probability 1 (with respect to the probability measure of $A_1, A_2, \cdots$).

If $\lambda \geq \frac{2}{m}$, then

$$\gamma_{j,m,t}^2 \leq \frac{16m\log m + 8m\log\lambda}{\lambda\nu t} \left(|\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i\in[m]} \inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right).$$

If $0 < \lambda < \frac{2}{m}$, then

$$\gamma_{j,m,t}^2 \leq \frac{4m^2}{\nu t} \left(|\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i\in[m]} \inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right).$$

Combining the upper bounds with Equation (B.6), we have:

If $\lambda \geq \frac{2}{m}$, then

$$\mathbb{P}_0 P_t^j d(\theta, \theta_0) \lesssim \frac{1}{t} + \frac{16m\log m + 8m\log\lambda}{\lambda\nu t} \left(|\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i\in[m]} \inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right) + \frac{1}{m^2 t} \sum_{j=1}^{m} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j).$$

If $0 < \lambda < \frac{2}{m}$, then

$$\mathbb{P}_0 P_t^j d(\theta, \theta_0) \lesssim \frac{1}{t} + \frac{4m^2}{\nu t} \left(|\mathbb{P}_0 \log \mathbb{p}_0| + \max_{i\in[m]} \inf_{\theta\in\Theta} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_\theta^i) \right) + \frac{1}{m^2 t} \sum_{j=1}^{m} D_{KL}(\mathbb{P}_0 \parallel \mathbb{P}_{\theta_0}^j).$$

This is the desired result after we remove the constants. ■

B.4 Proofs of Results in Section 8

Proof [Lemma 22] By Definition (2.9),

$$\log p_t^j(\theta) = t f_t^j(\theta) + \pi(\theta) + const.$$

for f_t^j defined in Equation (2.10).

We can compute $f_t^j(\theta)$ explicitly ,

$$\begin{aligned}
f_t^j(\theta) &= -\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\log \mathbb{p}_\theta^i(X_t^i) \\
&= -\frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\left[\langle\theta, T^i(X_k^i)\rangle - \psi^i(\theta)\right] + const \\
&= -\frac{1}{t}\left[\langle\theta, \sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]T^i(X_k^i)\rangle - \sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\psi^i(\theta)\right] + const.
\end{aligned}$$

This implies that

$$\begin{aligned}
\log p_t^j(\theta) &= -t f_t^j(\theta) + \pi(\theta) + const \\
&= \langle\theta, \sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]T^i(X_k^i)\rangle - \sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\psi^i(\theta) + \langle\theta, u\rangle - \psi^0(\theta) + const \\
&= \langle\theta, \sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]T^i(X_k^i) + u\rangle - \sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\psi^i(\theta) - \psi^0(\theta) + const.
\end{aligned}$$

This shows that p_t^j is a member of the exponential family of the form (8.3) with sufficient statistic $\chi_t^j + u$ and log-partition function $B_t^j(\theta) + \psi^0(\theta)$ as defined in Equation (8.4). ■

Proof [Lemma 23] By Lemma 22, P_t^j is a member of the exponential family given by Equation (8.3). The full rankness of P_t^j follows directly from the definition of full-rank exponential family and the compact representation (8.3). By Lemma 4.5 of Van der Vaart (2000), the log-partition function $B_t^j : \Theta \to R$ is infinitely times differentiable with the gradient being the mean parameter of χ_t^j and the Hessian is the covariance matrix of χ_t^j. Since P_t^j is full-rank, it is equivalent to the nonsingularity of the covariance matrix for χ_t^j Van der Vaart (2000). This implies that the gradient $\nabla_\theta B_t^j$ is one-to-one in the interior of Θ, thus, $\nabla_\theta B_t^j$ is invertible. ■

Proof [Proposition 24] By Lemma (23), the sequence of estimators $\hat{\theta}_t^j$ is well-defined with $[\mathbb{P}_0]-$probability tending to 1. By Theorem 4.6 of Van der Vaart (2000), the sequence of estimators $\hat{\theta}_t^j$ has limiting distribution

$$\sqrt{t}(\hat{\theta}_t^j - \theta_0) \xrightarrow{d} N(0, I_{\theta_0}^{j^{-1}}),$$

where $I_{\theta_0}^j$ is the Fisher information for $\mathbb{P}_\theta^j$ at θ_0. This corresponds to the Sandwich covariance matrix.

$$I_{\theta_0}^j = \left[\phi_{\theta_0}^{\prime -1}\mathrm{Cov}_{\theta_0}(\chi_t^j)\phi_{\theta_0}^{-1}\right]^{-1} = \mathrm{Cov}_{\theta_0}(\chi_t^j).$$

Since B_t^j is infinitely times differentiable in a neighborhood of θ_0, the Fisher information I_θ^j must be uniformly bounded. Thus, $\nabla \log \mathbb{p}_\theta^j$ is Lipchistz around θ_0 and Assumption 2(d) is satisfied.

When $\text{int}(\Theta) \neq \emptyset$, the exponential family $\mathbb{P}^i_\Theta$ satisfies the differential in the quadratic mean Assumption 2(c) by Example 7.7 of Van der Vaart (2000). Given that $\theta_0 \in \text{int}(\Theta)$ and π have support on Θ, the prior π puts positive mass on every neighborhood of θ_0, thus Assumption 3(b) is satisfied.

Under the full-rank assumptions on $\mathbb{P}^j_\theta$, the Hessian $\nabla^2_\theta \log \mathbb{p}^j_\theta$ is negative definite for all θ. Thus, the Hessian of f^j_t is negative definite as a positive linear combination of $\nabla^2_\theta \log \mathbb{p}^j_\theta$. We have that f^j_t is strictly concave on Θ with a unique maximizer at $\tilde{\theta}^j_t$. The strict concavity implies that

$$\lim_{t\to\infty} \mathbb{P}_0\Big(\inf_{\|\theta-\hat{\theta}^j_t\|>\delta} |f^j_t(\theta) - f^j_t(\hat{\theta}^j_t)| \geq \epsilon\Big) = 1$$

By Lemma 7, $\tilde{\theta}^j_t$ is consistent at θ_0. Then Assumption 4(a) is satisfied as a consequence of Lemma 8.

The assumptions of Theorem 10 are all satisfied. Considering the sequence of random variables $\theta \sim P^j_t$ centered at the moment estimator $\hat{\theta}^j_t$, we have

$$\int_\Theta |q^j_t(x) - N(0, V^{-1}_{\theta_0})| dx \xrightarrow{\mathbb{P}_0} 0,$$

where q^j_t is the density of $\sqrt{t}(\theta - \hat{\theta}^j_t)$, and $V_{\theta_0} = \frac{1}{m}\sum_{i=1}^m \text{Cov}(T^i)$. ■

Proof [Proposition 25] The prior π is assumed to have full support over $\mathbb{R}^p$, which trivially satisfies Assumption 3(a).

Define functions f^j_t, f_t, and f on Θ as per Equations (2.10)–(2.12). By Assumption (2)(a), the function f exists and we note that $f^j_t(\theta) \xrightarrow{\mathbb{P}_0} f(\theta)$.

The estimators $\hat{\theta}^j_t$ and the true parameter θ_0 are respectively given by:

$$\hat{\theta}^j_t = \arg\min_{\theta\in\Theta} f^j_t(\theta), \quad \theta_0 = \arg\min_{\theta\in\Theta} f(\theta).$$

The gradient of f^j_t is given by

$$\nabla f^j_t(\theta) = -\frac{1}{t} T^j_t + \frac{1}{t}\sum_{i=1}^m \sum_{k=1}^t \frac{[A^{t-k}_{ji}] X^j_k e^{\langle \theta, X^j_k\rangle}}{1 + e^{\langle \theta, X^j_k\rangle}},$$

and the Hessian is

$$\nabla^2 f^j_t(\theta) = \frac{1}{t}\sum_{i=1}^m \sum_{k=1}^t \frac{[A^{t-k}_{ji}] X^j_k X^{j^T}_k e^{\langle \theta, X^j_k\rangle}}{\left(1 + e^{\langle \theta, X^j_k\rangle}\right)^2}.$$

Given that $\nabla^2 f^j_t(\theta) > 0$, the function f^j_t is strictly convex in θ. This implies that the minimizer $\hat{\theta}^j_t$ is unique, and we have $\hat{\theta}^j_t \in \text{int}(\Theta)$.

The gradient of f is given by

$$\nabla f(\theta) = \mathbb{P}_{\theta_0}\left[\frac{X^1_1 e^{\langle \theta, X^1_1\rangle}}{1 + e^{\langle \theta, X^1_1\rangle}} - X^1_1 Y^1_1 \right] = \mathbb{P}_{\theta_0}\left[X^1_1 \left(\frac{e^{\langle \theta, X^1_1\rangle}}{1 + e^{\langle \theta, X^1_1\rangle}} - \frac{e^{\langle \theta_0, X^1_1\rangle}}{1 + e^{\langle \theta_0, X^1_1\rangle}} \right)\right],$$

The differentiation and expectation are interchanged by the Lebesgue-dominated convergence theorem since $|\nabla f(\theta)| \leq |X_1^1|$ and $\mathbb{P}_0|X_1^1| < \infty$ by Assumption ii).

Since $\nabla f(\theta)$ is strictly increasing in θ, we have $\nabla f(\theta - \epsilon) < 0 < \nabla f(\theta + \epsilon)$ in each coordinate. Hence, Assumption 2(b) is satisfied. By Lemma 7, we have $\hat{\theta}_t^j \xrightarrow{\mathbb{P}_0} \theta_0$.

Let $\mathcal{E} = \{\eta \in \mathbb{R} : |\sigma(\eta)| < \infty\}$. The set $\mathcal{E}$ is open and nonempty. Additionally, η is identifiable, as $\sigma(\eta)$ is a one-to-one function. Trivially, the set Θ is open and convex, and $\theta^T x \in \mathcal{E}$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$.

For all $\eta \in \mathcal{E}$, we have $0 < \sigma(\eta) < 1$ and

$$|\sigma'''(\eta)| = |\frac{\sigma(1-\sigma(\eta))(1-2\sigma(\eta))^2 - 2\sigma^2(1-\sigma(\eta))^2}{(1-\sigma(\eta))^2}| \leq 3.$$

After algebraic manipulations, we have

$$\nabla_\theta^3 \log \mathbb{p}_\theta^j(. \mid x_k^j)_{a,b,c} = \sigma'''(\langle\theta, x_k^j\rangle) x_{k,a}^j x_{k,b}^j x_{k,c}^j \leq 3 x_{k,a}^j x_{k,b}^j x_{k,c}^j,$$

for all $\theta \in \Theta, x_k^j \in \mathcal{X}$ and $a, b, c \in [p]$. Thus, $\nabla_\theta^3 \log \mathbb{p}_\theta^j$ is uniformly bounded by an integrable function.

The Fisher information for agent j is given by

$$V_{\theta_0}^j = -\mathbb{P}_0 \nabla_\theta^2 \log \mathbb{p}_\theta^j(. \mid X_k^j) = \mathbb{P}_0 \left[\frac{X_k^{j^T} e^{\langle\theta, X_k^j\rangle} X_k^j}{\left(1 + e^{\langle\theta, X_k^j\rangle}\right)^2} \right].$$

By Assumption i), $V_{\theta_0}^j$ exists and is nonsingular. Hence, Assumption 2(f) is satisfied.

Assumption 2(c), 4(a) are satisfied with the same argument as Proposition 24. By Corollary 12 and Slusky's theorem, we have

$$\int_\Theta \left| q_t^j(x) - N(0, \hat{V}_{\theta_0}^{-1}) \right| \, dx \xrightarrow{\mathbb{P}_0} 0.$$

where $\hat{V}_{\theta_0}$ is given by as the finite - sample version of the $\frac{1}{m}\sum_{j=1}^m V_{\theta_0}^j$. Specifically,

$$\hat{V}_\theta = \frac{1}{m} \sum_{i=1}^m X^{i^T} \text{diag} \left(\frac{e^{\sum_{j=0}^p \theta_j x_{1j}^i}}{\left(1 + e^{\sum_{j=0}^p \theta_j x_{1j}^i}\right)^2}, \ldots, \frac{e^{\sum_{j=0}^p \theta_j x_{tj}^i}}{\left(1 + e^{\sum_{j=0}^p \theta_j x_{tj}^i}\right)^2} \right) X^i.$$

$\blacksquare$

Proof [Lemma 26] By Lemma 2, $f_t^j(\theta) \xrightarrow{\mathbb{P}_0} f(\theta)$ for each $\theta \in \Theta$. For all $\theta \in [0,1]^2$, f_t^j is uniformly bounded.

$$\begin{aligned}
|f_t^j(\theta)| &\leq \frac{1}{t} \sum_{k=1}^t \sum_{i=1}^m [A_{ji}^{t-k}] \left[\frac{(|X_t^j - \theta|)^2}{2\sigma^{j^2}} + \log \left(\Phi(\frac{|Z^j| + \frac{1}{2}}{\sigma^j}) - \Phi(\frac{-\sqrt{2} - |Z^j|}{\sigma^j}) \right) \right] \\
&\leq \frac{1}{t} \sum_{k=1}^t \sum_{i=1}^m [A_{ji}^{t-k}] \left[\frac{1}{\sigma^{j^2}} + \log \left(\Phi(\frac{|Z^j| + \frac{1}{2}}{\sigma^j}) - \Phi(\frac{-\sqrt{2} - |Z^j|}{\sigma^j}) \right) \right] \\
&\leq \left[\frac{1}{\sigma^{j^2}} + \log \left(\Phi(\frac{|Z^j| + \frac{1}{2}}{\sigma^j}) - `\Phi(\frac{-\sqrt{2} - |Z^j|}{\sigma^j}) \right) \right] \left(1 + \frac{16 m^2 \log m}{\nu} \right).
\end{aligned}$$

The gradient $\nabla f_t^j(\theta)$ is uniformly bounded over $[0,1]^2$. To see this, consider

$$\nabla f_t^j(\theta) = A + B,$$

where A, B are given by

$$A = \frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\left[\frac{\theta - Z^j}{{\sigma^j}^2} - \frac{|X_t^j - Z^j|(\theta - Z^j)}{|\theta - Z^j|{\sigma^j}^2}\right],$$

$$B = \frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\frac{(\theta - Z^j)\left[\phi\left(\frac{|Z^j|+\frac{1}{2}-|\theta-Z^j|}{\sigma^j}\right) - \phi\left(\frac{-|\theta-Z^j|}{\sigma^j}\right)\right]}{\sigma^j|\theta - Z^j|\left[\Phi\left(\frac{|Z^j|+\frac{1}{2}-|\theta-Z^j|}{\sigma^j}\right) - \Phi\left(\frac{-|\theta-Z^j|}{\sigma^j}\right)\right]}.$$

Applying the following upper bounds on A and B,

$$A \leq \frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\frac{\sqrt{2}}{{\sigma^j}^2},$$

$$B \leq \frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]\frac{\phi\left(\frac{|Z^j|+\frac{1}{2}}{\sigma^j}\right) - \phi\left(\frac{-\sqrt{2}}{\sigma^j}\right)}{\sigma^j\left[\Phi\left(\frac{|Z^j|+\frac{1}{2}-\sqrt{2}}{\sigma^j}\right) - \Phi\left(\frac{-\sqrt{2}}{\sigma^j}\right)\right]},$$

we obtain

$$\|\nabla f_t^j(\theta)\| = \|A + B\| \leq \left[\frac{\sqrt{2}}{{\sigma^j}^2} + \frac{\phi(\frac{|Z^j|+\frac{1}{2}}{\sigma^j}) - \phi(\frac{-\sqrt{2}}{\sigma^j})}{\sigma^j\left[\Phi(\frac{|Z^j|+\frac{1}{2}-\sqrt{2}}{\sigma^j}) - \Phi(\frac{-\sqrt{2}}{\sigma^j})\right]}\right]\left(1 + \frac{16m^2\log m}{\nu}\right).$$

Then f_t^j is uniformly equicontinuous in t, as f_t^j is Lipschitz. By Lemma 36, we have $\|f_t^j - f\|_\infty \xrightarrow{\mathbb{P}_0} 0$.

Note that $\hat{\theta}_t^j = \arg\min_{\theta\in[0,1]^2} f_t^j(\theta)$ and $\theta_0 = \arg\min_{\theta\in[0,1]^2} f(\theta)$. By the Argmax theorem (Theorem 3.2.2, Vaart and Wellner (2023)), we have $\hat{\theta}_t^j \xrightarrow{\mathbb{P}_0} \theta_0$.

Since $\theta_0 \in \Theta$, for small enough ϵ, we have

$$\begin{aligned}\mathbb{P}_0(|\hat{\theta}_t^j - \theta_0| > \epsilon) &= \mathbb{P}_0(|\hat{\theta}_t^j - \theta_0| > \epsilon, \hat{\theta}_t^j \notin \Theta) + \mathbb{P}_0(|\hat{\theta}_t^j - \theta_0| > \epsilon, \hat{\theta}_t^j \in \Theta)\\ &= \mathbb{P}_0(\hat{\theta}_t^j \notin \Theta) + \mathbb{P}_0(|\hat{\theta}_t^j - \theta_0| > \epsilon, \hat{\theta}_t^j \in \Theta) \to 0,\end{aligned}$$

thus $\mathbb{P}_0(\hat{\theta}_t^j \notin \Theta) \to 0$. ■

Proof [Proposition 27] The prior π is assumed to have full support over Θ, which trivially satisfies Assumption 3(a).

From the proof of Lemma 26, we proved that ∇f_t^j is uniformly bounded for all t, thus f_t^j is uniformly equicontinuous. By definition, there exists $\delta > 0$ such that $\|\theta - \hat{\theta}_t^j\| > \delta$ implies that $|f_t^j(\theta) - f_t^j(\hat{\theta}_t^j)| > \epsilon$. This verifies Assumption 4(a).

The gradient of $\log \mathbb{p}_\theta^j$ is given by

$$\nabla \log \mathbb{p}_\theta^j(X_t^j) = \frac{\theta - Z^j}{|\theta - Z^j|{\sigma^j}^2}\left[|\theta - Z^j| - |X_t^j - Z^j| - \frac{\phi\left(\frac{|Z^j|+\frac{1}{2}-|\theta-Z^j|}{\sigma^j}\right) - \phi\left(\frac{-|\theta-Z^j|}{\sigma^j}\right)}{\Phi\left(\frac{|Z^j|+\frac{1}{2}-|\theta-Z^j|}{\sigma^j}\right) - \Phi\left(\frac{-|\theta-Z^j|}{\sigma^j}\right)}\right].$$

This allows us to derive the Fisher information of p_θ^j.

$$\begin{aligned} V_\theta^j &= \mathbb{P}_0[\nabla_\theta \log \mathbb{p}_\theta^j \nabla_\theta \log {\mathbb{p}_\theta^j}^T] \\ &= \frac{(\theta - Z^j)(\theta - Z^j)^T}{{\sigma^j}^4|\theta - Z^j|^2}\left[|\theta - Z^j| - |\theta_0 - Z^j| - \frac{\phi(\frac{|Z^j|+\frac{1}{2}-|\theta-Z^j|}{\sigma^j}) - \phi(\frac{-|\theta-Z^j|}{\sigma^j})}{\Phi(\frac{|Z^j|+\frac{1}{2}-|\theta-Z^j|}{\sigma^j}) - \Phi(\frac{-|\theta-Z^j|}{\sigma^j})}\right]^2. \end{aligned}$$

At θ_0, the Hessian simplifies to $V_{\theta_0}^j = \frac{(\theta_0-Z^j)(\theta_0-Z^j)^T}{{\sigma^j}^4|\theta_0-Z^j|^2}\left[\frac{\phi(\frac{|Z^j|+\frac{1}{2}-|\theta_0-Z^j|}{\sigma^j})-\phi(\frac{-|\theta_0-Z^j|}{\sigma^j})}{\Phi(\frac{|Z^j|+\frac{1}{2}-|\theta_0-Z^j|}{\sigma^j})-\Phi(\frac{-|\theta_0-Z^j|}{\sigma^j})}\right]^2$ which is nonsingular. Thus, Assumption 2(e) is satisfied.

Reusing the argument in the proof of Lemma 26, we have

$$\sup_{\theta\in[0,1]^2} \|\nabla \log p_\theta^j\| \le \frac{\sqrt{2}}{{\sigma^j}^2} + \frac{\phi(\frac{|Z^j|+\frac{1}{2}}{\sigma^j}) - \phi(\frac{-\sqrt{2}}{\sigma^j})}{\sigma^j\left[\Phi(\frac{|Z^j|+\frac{1}{2}-\sqrt{2}}{\sigma^j}) - \Phi(\frac{-\sqrt{2}}{\sigma^j})\right]}.$$

This implies that $\nabla \log \mathbb{p}_\theta^j$ is a continuously differentiable function over a compact domain $[0,1]^2$. Thus, it is Lipschitz continuous. Assumption 2(d) is satisfied.

The Hessian for f_t^j is simply the average of agent Fisher information:

$$\nabla^2 f_t^j(\theta) = \frac{1}{t}\sum_{k=1}^{t}\sum_{i=1}^{m}[A_{ji}^{t-k}]V_\theta^i.$$

Recall that in the proof of Since f_t^j is uniformly equicontinuous,

We define the event E_t as

$$E = \{X^{(mt)} : \hat{\theta}_t^j \in \Theta\}.$$

Conditioning on E_t, by Corollary 12 and Slusky's theorem, we have

$$\int_\Theta \left|q_t^j(x) - N(0, V_{\theta_0}^{-1})\right| dx \xrightarrow{\mathbb{P}_0} 0,$$

for q_t^j defined as the density of $\sqrt{t}(\theta - \hat{\theta}_t^j)$ and V_{θ_0} defined in Equation (8.16).

By Lemma 26, $\lim_{t\to\infty} \mathbb{P}_0(E_t) = 1$. This completes our proof. ■

Appendix C. Supporting Results

Proof [Lemma 1] Define $\delta = \max(|\lambda_2(A)|, |\lambda_m(A)|)$. The spectral radius of A is 1, thus $\delta < 1$ by Perron-Frobenius theorem. Under the assumptions 1, by the standard property of stochastic matrices (see e.g. Rosenthal (1995)), the diagonalizable matrix A satisfies

$$\|e_i^T A^t - \frac{1}{m}\mathbf{1}\|_1 \leq m\delta^t \tag{B.7}$$

for any $i \in [m]$, using the fact that $\frac{1}{m}\mathbf{1}$ is the stationary distribution of the Markov chain with transition matrix A.

For any $t - k \geq \tilde{t} = \frac{\log \frac{m}{2}}{-\log \delta}$,

$$m\delta^{t-k} \leq m\delta^{\tilde{t}} \leq \frac{m}{\frac{m}{2}} \leq 2$$

Since $\|e_i^T A^t - \frac{1}{m}\mathbf{1}\|_1 \leq 2$ by the double stochasticity of A, we use (B.7) to break the quantity of interest $\sum_{k=1}^t \sum_{j=1}^m |[A_{ij}^{t-k}] - \frac{1}{m}|$ into two parts, that is, for any $i \in [m]$,

$$\begin{aligned}
\sum_{k=1}^{t}\sum_{j=1}^{m} |[A_{ij}^{t-k}] - \frac{1}{m}| &= \sum_{k=1}^{t} \|e_i^T A^{t-k} - \frac{1}{m}\mathbf{1}\|_1 \\
&= \sum_{k=1}^{t-\tilde{t}} \|e_i^T A^{t-k} - \frac{1}{m}\mathbf{1}\|_1 + \sum_{k>t-\tilde{t}}^{t} \|e_i^T A^{t-k} - \frac{1}{m}\mathbf{1}\|_1 \\
&\leq \sum_{k=1}^{t-\tilde{t}} m\delta^{t-k} + 2\tilde{t} \leq \frac{m\delta^{\tilde{t}}}{1-\delta} + 2\tilde{t} \leq \frac{2}{1-\delta} + \frac{2\log\frac{m}{2}}{-\log\delta}
\end{aligned}$$

Noting that $1 - \delta \leq -\log\delta$ and $\delta = 1 - \frac{\nu}{4m^2}$ for any doubly stochastic matrix A, we get

$$\sum_{k=1}^{t}\sum_{j=1}^{m} |[A_{ij}^{t-k}] - \frac{1}{m}| \leq \frac{2 + 2\log\frac{m}{2}}{1-\delta} \leq \frac{4\log m}{1-\delta} \leq \frac{16m^2\log m}{\nu}$$

■

Lemma 34 (Distributed Law of Large Numbers) *Assume that $S_t^i, t \geq 1$ are i.i.d. random variables with $\mathbb{E}[|S_t^i|]$ exists and are finite, for all $i \in [m]$. Under Assumption 1, for any $j \in [m]$, the random variables*

$$Z_t^j = \frac{1}{t}\sum_{i=1}^{m}\sum_{k=1}^{t}[A_{ij}^{t-k}]S_t^i$$

converge in probability to $\frac{1}{m}\sum_{i=1}^m \mathbb{E}[S_1^i]$ as $t \to \infty$.

Proof Let $Z_t = [Z_t^1, \cdots, Z_t^m]^T$ and $S_t = [S_t^1, \cdots, S_t^m]^T$. Since $\mathbb{E}[|S_t^i|] < \infty$ exists and is finite, we have

$$Z_t = \frac{1}{t}\sum_{k=1}^{t} A^{t-k} S_k.$$

Using the property of stochastic matrices (Rosenthal, 1995), the diagonalizable matrix A satisfies

$$\|e_i^T A^t - \frac{1}{m}\mathbf{1}\|_1 \le m\delta^t.$$

for all $i \in [m]$ and some $\delta < 1$.

This implies hat

$$\begin{aligned}
&\left\| Z_t - \sum_{k=1}^{t} \frac{1}{m}\mathbf{1}\mathbf{1}^T S_k \right\| = \left\| \frac{1}{t}\sum_{k=1}^{t} A^{t-k} S_k - \frac{1}{m}\mathbf{1}\mathbf{1}^T S_k \right\| \\
&\le \frac{1}{t}\sum_{k=1}^{t} \|(A^{t-k} - \frac{1}{m}\mathbf{1}\mathbf{1}^T) S_k\| \\
&\le \frac{1}{t}\sum_{k=1}^{t} \max_{i\in[m]} \left\| e_i^T A^{t-k} - \frac{1}{m}\mathbf{1} \right\| \max_{i\in[m]} |S_k^i|, \quad \text{by Cauchy - Schwarz inequality} \\
&\le \frac{1}{t}\sum_{k=1}^{t} m\delta^{t-k} \max_{i\in[m]} |S_k^i| \\
&\le \frac{m}{t}\sum_{k=1}^{t}\sum_{i=1}^{m} \delta^{t-k}|S_k^i| \\
&\le \frac{m}{t}\sum_{k=t-c+1}^{t}\sum_{i=1}^{m} \delta^{t-k}|S_k^i| + \frac{m}{t}\delta^c \sum_{k=1}^{t-c}\sum_{i=1}^{m} |S_k^i|,
\end{aligned}$$

for some fixed constant c.

By the strong law of large number, $\left\|\sum_{k=1}^{t} \frac{1}{m}\mathbf{1}\mathbf{1}^T S_k - \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[S_1^i]\right\| = o_d(1)$. Then

$$\begin{aligned}
\left\| Z_t - \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[S_1^i] \right\| &\le \left\| Z_t - \sum_{k=1}^{t} \frac{1}{m}\mathbf{1}\mathbf{1}^T S_k \right\| + \left\| \sum_{k=1}^{t} \frac{1}{m}\mathbf{1}\mathbf{1}^T S_k - \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[S_1^i] \right\| \\
&= \frac{m}{t}\delta^c \sum_{k=1}^{t-c}\sum_{i=1}^{m} |S_k^i| + \frac{m}{t}\sum_{k=t-c+1}^{t}\sum_{i=1}^{m} \delta^{t-k}|S_k^i| + o_d(1) \\
&\le \frac{m}{t}\delta^c \sum_{k=1}^{t-c}\sum_{i=1}^{m} |S_k^i| + o_d(1) \\
&\le m\delta^c \sum_{i=1}^{m} \mathbb{E}[S_1^i] + o_d(1).
\end{aligned}$$

Since this for arbitrarily large c, we conclude that $\left|Z_t - \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[S_1^i]\right| = o_d(1)$, that is

$$Z_t \to \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}[S_1^i] \quad \text{in probability.}$$

■

Lemma 35 (Distributed Central Limit Theorem) *Assume that $S_t^i, t \geq 1$ are i.i.d. random variables with $\mathbb{E}[S_t^i] = \mu^i$ and $cov[S_t^i] = \Sigma^i$ exists and are finite, for all $i \in [m]$. Under Assumption 1, for any $j \in [m]$, the random variables*

$$Z_t^j = \sqrt{\frac{m}{t}} \left(\frac{1}{m} \sum_{i=1}^m \Sigma^i \right)^{-1/2} \sum_{i=1}^m \sum_{k=1}^t [A_{ij}^{t-k}](S_t^i - \mu^i)$$

converge in distribution to a standard normal distribution as $t \to \infty$.

Proof The proof applies the Linderberg central limit theorem. Without loss of generality, assume that $\mu^i = 0$. Let $\Sigma_{m,t}$ be the sum of covariance. We have

$$\Sigma_{m,t} = \sum_{i=1}^m \sum_{k=1}^t \mathrm{Cov}([A_{ij}^{t-k}] S_t^i) = \sum_{i=1}^m \sum_{k=1}^t [A_{ij}^{t-k}]^2 \Sigma^i.$$

The matrix $\Sigma_{m,t}$ is asymptotically equivalent to the scaling factor $\Sigma_{m,t}^* = \frac{t}{m^2} \sum_{i=1}^m \Sigma^i$.

$$\begin{aligned}
\left(\Sigma_{m,t}^*\right)^{-1} \Sigma_{m,t} &= \frac{m}{t} \left(\frac{1}{m} \sum_{i=1}^m \Sigma^i \right)^{-1} \sum_{i=1}^m \sum_{k=1}^t [A_{ij}^{t-k}]^2 \Sigma^i \\
&= m \left(\frac{1}{m} \sum_{i=1}^m \Sigma^i \right)^{-1} \sum_{i=1}^m \Sigma^i \frac{\sum_{k=1}^t [A_{ij}^{t-k}]^2}{t} \\
&= m \left(\frac{1}{m} \sum_{i=1}^m \Sigma^i \right)^{-1} \sum_{i=1}^m \Sigma^i \frac{\sum_{k=1}^t [A_{ij}^{k}]^2}{t}
\end{aligned}$$

By Assumption 1, $\lim_{t\to\infty} A_{ij}^k \to \frac{1}{m}$, thus $\lim_{t\to\infty} [A_{ij}^k]^2 \to \frac{1}{m^2}$.Then we have

$$\begin{aligned}
\lim_{t\to\infty} \left(\Sigma_{m,t}^*\right)^{-1} \Sigma_{m,t} &= m \left(\frac{1}{m} \sum_{i=1}^m \Sigma^i \right)^{-1} \sum_{i=1}^m \Sigma^i \lim_{t\to\infty} \frac{\sum_{k=1}^t [A_{ij}^{k}]^2}{t} \\
&= m \left(\frac{1}{m} \sum_{i=1}^m \Sigma^i \right)^{-1} \sum_{i=1}^m \Sigma^i \frac{1}{m^2} = I_m.
\end{aligned}$$

It remains to verify the Linderberg condition

$$\lim_{t\to\infty} \sum_{i=1}^m \sum_{k=1}^t \mathbb{E}\left[\|\Sigma_{m,t}^{-1/2} [A_{ij}^{t-k}] S_t^i\|^2 I_{\|\Sigma_{m,t}^{-1/2} [A_{ij}^{t-k}] S_t^i\| > \epsilon} \right] = 0, \quad \forall \epsilon > 0.$$

The condition is equivalent to

$$\lim_{t\to\infty} \sum_{i=1}^m \sum_{k=1}^t \mathbb{E}\left[\|\Sigma_{m,t}^{*-1/2} [A_{ij}^{t-k}] S_t^i\|^2 I_{\|\Sigma_{m,t}^{-1/2} [A_{ij}^{t-k}] S_t^i\| > \epsilon} \right] = 0, \quad \forall \epsilon > 0.$$

Since $\lim_{t\to\infty} \Sigma^*_{m,t} = \infty$, $\lim_{t\to\infty} \Sigma_{m,t} = \infty$ and $I_{\|\Sigma_{m,t}^{-1/2}[A_{ij}^{t-k}]S_t^i\|>\epsilon}$ converges to 0 almost surely.

Because $\max_{i\in[m]} \sup_t [A^t_{ij}]^2 < \infty$, we have

$$\begin{aligned}\max_{i\in[m]}\sup_t t\mathbb{E}\left[\|\Sigma_{m,t}^{*^{-1/2}}[A_{ij}^{t-k}]S_t^i\|^2\right] &= \max_{i\in[m]}\sup_t \mathbb{E}\left[\|m(\frac{1}{m}\sum_{i=1}^m \Sigma^i)^{-1}[A_{ij}^{t-k}]S_t^i\|^2\right] \\ &\leq m(\max_{i\in[m]}\sup_t [A^t_{ij}]^2)\|(\frac{1}{m}\sum_{i=1}^m \Sigma^i)^{-1}\|_{op} \max_{i\in[m]} \|\Sigma^i\|_{op}\end{aligned}$$

By Lebesgue dominated convergence theorem, we have

$$\lim_{t\to\infty} t\mathbb{E}\left[\|\Sigma_{m,t}^{*^{-1/2}}[A_{ij}^{t-k}]S_t^i\|^2 I_{\|\Sigma_{m,t}^{-1/2}[A_{ij}^{t-k}]S_t^i\|>\epsilon}\right] \to 0.$$

The Cesaro sums also converge to 0:

$$\lim_{t\to\infty}\sum_{i=1}^m \frac{1}{t}\sum_{k=1}^t t\mathbb{E}\left[\|\Sigma_{m,t}^{*^{-1/2}}[A_{ij}^{t-k}]S_t^i\|^2 I_{\|\Sigma_{m,t}^{-1/2}[A_{ij}^{t-k}]S_t^i\|>\epsilon}\right] = 0, \quad \forall \epsilon > 0.$$

This completes the proof ■

Lemma 36 (Miller (2021)) *Suppose that $h_n : E \to F$ for $n \in \mathbb{N}$, where E is a totally bounded space and F is a normed space. If h_n converges pointwise and is equicontinuous, then it converges uniformly.*